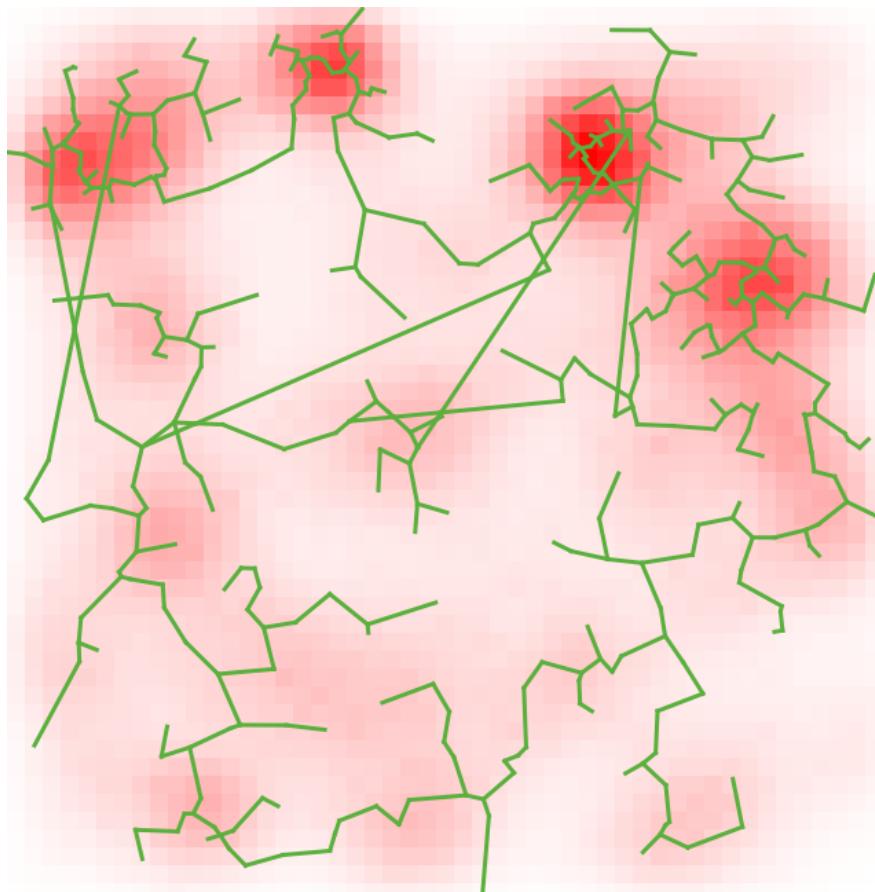


TOWARDS MODELS COUPLING URBAN GROWTH AND TRANSPORTATION NETWORK GROWTH

JUSTE RAIMBAULT



PhD Thesis Second Year Preliminary Memoire

Under the supervision of ARNAUD BANOS and FLORENT LE NÉCHET

UMR CNRS 8504 Géographie-cités
and UMR-T IFSTTAR 9403 LVMT

Université Paris VII

February 2017 – version 1.1

Juste Raimbault : *Towards Models Coupling Urban Growth and Transportation Network Growth*, PhD Thesis Second Year Preliminary Memoire,
© February 2017

ABSTRACT

READING NOTES

This provisory Memoire must be read as a work in progress, as it details progresses after one year of Doctorate. Many parts are given at the state of project, and not omitted as playing a role in the current research questioning. Its purpose is to set up a plan and examine the achieved work and corresponding directions, but also to share research ideas at this important step of one year.

PUBLICATIONS

Les travaux suivants contiennent une grande partie du contenu de cette thèse :

PUBLICATIONS

Antelope, C., Hubatsch, L., Raimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016.*

Raimbault, J. (2017). A Discrepancy-Based Framework to Compare Robustness Between Multi-attribute Evaluations. In *Complex Systems Design & Management* (pp. 141-154). Springer International Publishing.

Raimbault, J. (2016). Investigating the Empirical Existence of Static User Equilibrium, *forthcoming in EWGT 2016 proceedings, Transportation Research Procedia*. arxiv :1608.05266

Raimbault, J. (2016). Generation of Correlated Synthetic Data, forthcoming in *Actes des Journées de Rochebrune 2016*.

Raimbault, J. (2015). Models Coupling Urban Growth and Transportation Network Growth : An Algorithmic Systematic Review Approach, forthcoming in *ECTQG 2015 proceedings*. arxiv :1605.08888

COMMUNICATIONS

Towards a Theory of Co-evolutive Networked Territorial Systems : Insights from Transportation Governance Modeling in Pearl River Delta, China, *MEDIUM Seminar : Sustainable Development in Zhuhai, Guangzhou, Dec 2016*.

Models of growth for system of cities : Back to the simple, *Conference on Complex Systems 2016, Amsterdam, Sep 2016*.

For a Cautious Use of Big Data and Computation. *Royal Geographical Society - Annual Conference 2016 - Session : Geocomputation, the Next 20 Years (1), London, Aug 2016*.

Indirect Bibliometrics by Complex Network Analysis. *20e Anniversaire de Cybergeo, Paris, May 2016*.

Raimbault, J. & Serra, H. (2016). Game-based Tools as Media to Transmit Freshwater Ecology Concepts, *poster corner at SETAC 2016 (Nantes, May 2016)*.

Le Néchet, F. & Raimbault, J. (2015). Modeling the emergence of metropolitan transport authority in a polycentric urban region, *ECTQG 2015, Bari, Sep 2015*.

Hybrid Modeling of a Bike-Sharing Transportation System, *poster presented at ICCSS 2015, Helsinki, June 2015*.

Raimbault, J. & Gonzales, J. (2015). Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d'Infrastructures de Transport, *poster presented at Rencontres du Labex Dynamite, Paris, May 2015*.

TABLE DES MATIÈRES

I FOUNDATIONS	11
1 INTERACTIONS BETWEEN NETWORKS AND TERRITORIES	13
1.1 Réseaux et Territoires	14
1.2 Modéliser les Interactions	20
1.3 -NoValue-	25
1.4 Question de Recherche	25
2 METHODOLOGICAL DEVELOPMENTS	27
2.1 Reproductibilité	28
2.2 Un cadre uniifié pour les modèles stochastiques de croissance urbaine	32
2.3 Sensibilité des Lois d'Echelle Urbaines à l'Etendue Spatiale	35
2.4 Contrôle statistique pour les conditions initiales par génération de données synthétiques	37
2.5 -NoValue-	39
2.6 -NoValue-	42
2.7 -NoValue-	45
2.8 Un Cadre basé sur la Discrépance pour Comparer la Robustesse des Evaluations Multi-attributs	48
3 QUANTITATIVE EPISTEMOLOGY	63
3.1 Revue Systématique Algorithmique	64
3.2 Bibliométrie Indirecte par Analyse de Réseaux Complexes	70
3.3 Vers une modélisation des thèmes et une extraction automatique du contexte	71
II MATERIALS	73
4 EMPIRICAL ANALYSIS : INSIGHTS FROM STYLIZED FACTS	75
4.1 Corrélations Statiques entre Forme Urbaine et Forme de Réseau	76
4.2 -NoValue-	84
4.3 -NoValue-	88
4.4 -NoValue-	90
4.5 Investigation Empirique de l'Existence de l'Equilibre Utilisateur Statique	91
5 MODELING	105
5.1 Un modèle simple de croissance urbaine	106
5.2 Génération de configurations territoriales corrélées . . .	107
5.3 -NoValue-	117
6 TRANSPORTATION GOVERNANCE MODELING	119
6.1 -NoValue-	119

III SYNTHESIS	121
7 A ROADMAP FOR AN OPERATIONAL FAMILY OF MODELS OF COEVOLUTION	123
7.1 Objectifs	123
7.2 Cas d'étude	123
7.3 Feuille de Route	124
IV OPENING	125
8 THEORETICAL FRAMEWORK	127
8.1 Pour une Théorie Géographique	128
8.2 Un Cadre Théorique pour l'Etude des Systèmes Sociaux-techniques	134
9 THEMATIC AND GENERAL PERSPECTIVES	145
9.1 Développement Spécifiques	146
9.2 Vers un Programme de Recherche	147
BIBLIOGRAPHIE	153
V APPENDIX	177
10 AN INTERDISCIPLINARY APPROACH TO MORPHOGENESIS	179
11 TECHNICAL DEVELOPMENTS	181
11.1 Dérivations pour les modèles de croissance urbaine	181
11.2 Sensibilité des Lois d'Echelle Urbaines	183
12 GENERATION OF CORRELATED SYNTHETIC DATA	187
13 QUANTITATIVE ANALYSIS OF THESIS REFLEXIVITY	193
14 DATASETS	195
14.1 Données de Traffic du Grand Paris	195
14.2 Réseau Routier Européen	195
14.3 Réseau Dynamique des Autoroutes Françaises	195
15 SOFTWARES AND PACKAGES	197
15.1 largeNetwoRk : Import de réseau et simplification pour R	197
15.2 Fouille de Corpus scientifique	197
16 ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION	199
16.1 Revue Systématique Algorithmique	199
16.2 Bibliométrie Indirecte	200
16.3 Croissance Urbaine	200
16.4 Génération des Données Synthétiques Corrélées	201
16.5 Modèle Lutecia	201
16.6 Analyse des Réseaux	201
17 TOOLS AND WORKFLOW FOR AN OPEN REPRODUCIBLE RESEARCH	203
17.1 Générateur de Documentation Netlogo	203
17.2 git comme outil de reproductibilité	203
17.3 git-data	203
17.4 Vers un gestionnaire de métadonnées compatible avec git	204

17.5 TorPool	204
-------------------------------	------------

TABLE DES FIGURES

FIGURE 1	Reproductibilité et visualisation	30
FIGURE 2	Cartes de ségrégation métropolitaine	58
FIGURE 3	Sensibilité de la robustesse aux données manquantes	60
FIGURE 4	-NoValue-	67
FIGURE 5	-NoValue-	68
FIGURE 6	-NoValue-	77
FIGURE 7	-NoValue-	78
FIGURE 8	-NoValue-	79
FIGURE 9	-NoValue-	89
FIGURE 10	Capture de l'application web permettant l'exploration spatio-temporelle des données de trafic pour la région Parisienne. Il est possible de choisir date et heure (précision de 15min sur un mois, réduite par rapport au jeu de données initial pour des raisons de performance). Un graphe résume les motifs de congestion pour la journée courante.	95
FIGURE 11	Variabilité spatiale d'un plus court chemin en temps de trajet (trajet du plus court chemin en pointillé bleu). Dans un intervalle de seulement 10 minutes, entre le 11/02/2016 00 :06 (à gauche) et le 11/02/2016 00 :16 (à droite), le plus court chemin entre Porte d'Auteuil à l'ouest et Porte de Bagnolet à l'est, augmente en distance effective de $\simeq 37\text{km}$ (avec une augmentation du temps de trajet de seulement 6 minutes), à cause d'une forte perturbation sur le périphérique parisien.	96
FIGURE 12	Variabilité maximale du temps de trajet (en haut en minutes et de la distance de trajet correspondante (en bas) pour un échantillon de deux semaines. Le graphe représente le maximum sur l'ensemble des paires Origine-Destination de la variabilité absolue entre deux pas de temps consécutifs. Les heures de pointe induisent une forte variabilité du temps de trajet, allant jusqu'à 25 minutes et une variabilité de distance jusqu'à 35km.	97

FIGURE 13	Stabilité temporelle du maximum de la centralité de chemin. Le graphe montre dans le temps la dérivée normalisée du maximum de la centralité de chemin, qui capture ses variations relatives à chaque pas de temps. La valeur maximale de 25% correspond à de très fortes perturbations du réseau sur les liens correspondants, puisque cela implique qu'au moins cette proportion d'utilisateurs prenant le lien dans des conditions précédentes doivent prendre un trajet complètement différent.	98
FIGURE 14	Auto-corrélations spatiales pour les vitesses relatives sur deux semaines. Le graphe montre les valeurs de l'auto-corrélation dans le temps, pour des valeurs variables (1,10km) de la distance de décroissance. les valeurs intermédiaires de la distance de décroissance donnent une déformation relativement continue entre ces deux extrêmes. Les points sont lissés sur une fenêtre temporelle de 2h pour faciliter la lecture. Les lignes pointillées verticales correspondent à minuit de chaque jour. La courbe violette donne la vitesse relative, ajustée à l'échelle pour établir la correspondance entre les heures de pointe et les variations de l'auto-corrélation.	100
FIGURE 15	-NoValue-	111
FIGURE 16	-NoValue-	112
FIGURE 17	-NoValue-	117
FIGURE 18	-NoValue-	185
FIGURE 19	-NoValue-	189
FIGURE 20	-NoValue-	191
FIGURE 21	-NoValue-	192

LISTE DES TABLEAUX

TABLE 1	Résultats numériques des simulations pour chaque arrondissement avec N = 50 répétitions. Chaque valeur des indicateurs factice est donnée par sa moyenne sur les répétitions et la déviation standard associée. Le ratio de robustesse est calculé par rapport au premier arrondissement (choix arbitraire). Un ratio inférieur à 1 signifie que la borne de l'intégrale est plus petite pour le premier système, i.e. que l'évaluation est plus robuste pour celui-ci.	57
TABLE 2	-NoValue-	68

INTRODUCTION

C'est quand on donne un coup de pied dans la fourmilière qu'on se rend compte de toute sa complexité.

- ARNAUD BANOS

"En conséquence d'un problème technique, le trafic est interrompu sur la ligne B du RER pour une durée indéterminée. Plus d'information seront fournies dès que possible". Il y a des fortes chances pour que quiconque ayant vécu ou passé un peu de temps en région parisienne ait déjà entendu cette annonce glaçante et en ait subi les conséquences pour le reste de la journée. Mais il ne se doute sûrement pas des ramifications des cascades causales induites par cet évènement presque banal. Les systèmes territoriaux, quelles que soient les aspects considérés pour leur définition, seront toujours extrêmement complexes, les interrelations à de nombreuses échelles spatiales et temporelles participant à la production des comportements émergents observés à tout niveau du système. Martin est un étudiant qui fait l'aller-retour journalier entre Paris et Palaiseau and manquera un examen crucial, ce qui aura un impact profond sur sa vie professionnelle : implications à une longue échelle de temps, une petite échelle spatiale et à la granularité de l'agent. Yuangsi était en train de relier les aéroports d'Orly et Roissy dans son voyage de Londres à Pékin et va manquer son avion ainsi que le mariage de sa soeur : grande échelle spatiale, petite échelle de temps, granularité de l'agent. Une pétition collective émerge des voyageurs, conduisant à la création d'une organisation qui mettra la pression sur les autorités pour qu'elles augmentent le niveau de service : échelle temporelle et spatiales mesoscopique, granularité de l'aggregation d'agents. La recherche de cause possible à l'incident conduira à des processus intriqués à diverses échelles, parmi lesquels aucun ne semble être une meilleure explication ; le développement historique du réseau ferroviaire en région parisienne a conditionné les évolutions futures et le RER B a suivi l'ancienne Ligne de Sceaux, le plan de DELOUVRIER pour le développement régional et son execution partielle, sont également des éléments d'explication des faiblesses structurelles du réseau parisien de transports en commun [111] ; le motifs pendulaires dus à l'organisation territoriale induisent une surcharge de certaines ligne et ainsi nécessairement une augmentation des incidents d'exploitation. La liste pourrait être ainsi continuée un certain temps, chaque approche apportant sa vision mature correspondant à un corpus de

connaissances scientifiques dans des disciplines diverses comme la géographie, l'économie urbaine, les transports. Cette anecdote amusante est suffisante pour faire ressentir la complexité des systèmes territoriaux. Notre but ici est de se plonger dans cette complexité, et en particulier donner un point de vue original sur l'étude des relations entre réseaux et territoires. Le choix de cette position sera largement discuté dans une partie thématique, nous nous concentrerons à présent sur l'originalité du point de vue que nous allons prendre.

DE LA POSITION GÉNÉRALE

L'ambition de cette thèse est de ne pas avoir d'ambition. Cette entrée en matière, rude en apparence, contient à différents niveaux les logiques sous-jacentes à notre processus de recherche. Au sens propre, nous nous plaçons tant que possible dans une démarche constructive et exploratoire, autant sur les plans théoriques et méthodologiques que thématique, mais encore proto-méthodologique (outils appliquant la méthode) : si des ambitions unidimensionnelles ou intégrées devaient émerger, elles seraient conditionnées par l'arbitraire choix d'un échantillon temporel parmi la continuité de la dynamique qui structure tout projet de recherche. Au sens structurel, l'auto-référence qui soulève une contradiction apparente met en exergue l'aspect central de la réflexivité dans notre démarche constructive, autant au sens de la récursivité des appareils théoriques, de celui de l'application des outils et méthodes développés au travail lui-même ou que de celui de la co-construction des différentes approches et des différents axes thématiques. Le processus de production de connaissance pourra ainsi être lu comme une métaphore des processus étudiés. Enfin, sur un plan plus enclin à l'interprétation, cela suggérera la volonté d'une position délicate liant un positionnement politique dont la nécessité est intrinsèque aux sciences humaines (par exemple ici contre l'application technocratique des modèles, ou pour le développement d'outils luttant pour une science ouverte) à une rigueur d'objectivité plus propre aux autres champs abordés, position forçant à une prudence accrue.

CONTEXTE SCIENTIFIQUE : PARADIGMES DE LA COMPLEXITÉ

Pour une meilleure introduction du sujet, il est nécessaire d'insister sur le cadre scientifique dans lequel nous nous positionnons. Ce contexte est crucial à la fois pour comprendre les concepts épistémologiques implicites dans nos questions de recherche, et aussi pour être conscient de la variété de méthodes et outils utilisés. La science contemporaine prend progressivement le tournant de la complexité dans de nombreux champs, ce qui implique une mutation épistémologique pour abandonner le réductionnisme strict qui a échoué

dans la majorité de ses tentatives de synthèse [7]. Arthur a rappelé récemment [9] qu'une mutation des méthodes et paradigmes en était également un enjeu, de par la place grandissante prise par les approches computationnelles qui remplacent les résolutions purement analytiques généralement limité en possibilités de modélisation et de résolution. La capture des *propriétés émergentes* par des modèles de systèmes complexes est une des façons d'interpréter la philosophie de ces approches.

Ces considérations sont bien connues des Sciences Humaines (qualitatives et quantitatives) pour lesquelles la complexité des agents et systèmes étudiés est une des justifications de leur existence : si les humains étaient des particules, la majorité des disciplines les prenant comme objet d'étude n'auraient jamais émergé puisque la thermodynamique aurait alors résolu la majorité des problèmes sociaux¹. Elles sont au contraire moins connues et acceptées en sciences "dures" comme la physique : LAUGHLIN développe dans [151] une vision de la discipline à la même position de "frontière des connaissances" que d'autre champs pouvant paraître moins matures. La plupart des connaissances actuelles concerne des structures classiques simples, alors qu'un grand nombre de système présentent des propriétés *d'auto-organisation*, au sens où les lois macroscopiques ne sont pas suffisantes pour inférer les propriétés macroscopiques du systèmes à moins que son évolution soit entièrement simulée (plus précisément cette vision peut être prise comme une définition de l'émergence sur laquelle nous reviendrons par la suite, or des propriétés auto-organisées sont par nature émergentes). Cela correspond au premier cauchemar du Démon de Laplace développé dans [84].

A la croisée de positionnements épistémologiques, de méthodes et de champs d'application, les *Sciences de la complexité* se concentrent sur l'importance de l'émergence et de l'auto-organisation dans la plupart des phénomènes réel, ce qui les place plus proche de la frontière des connaissances que ce que l'on peut penser pour des disciplines classiques (LAUGHLIN, op. cit.). Ces concepts ne sont pas récents et avaient déjà été mis en valeur par ANDERSON [7]. On peut aussi interpréter la Cybernétique comme un précurseur des Sciences de la Complexité en la lisant comme un pont entre technologie et sciences cognitives [263]. Plus tard, la Synergétique [123] a posé les bases d'approches théoriques des phénomènes collectifs en physique. Les causes possibles de la croissance récente du nombre de travaux se réclamant d'approches complexes sont nombreuses. L'explosion de la puissance de calcul en est certainement une vu le rôle central que jouent les simulations numériques [251]. Elles peuvent aussi être à chercher auprès de progrès en épistémologie : introduction

¹ bien que cette affirmation soit elle-même discutable, les sciences physiques classiques ayant également échoué à prendre en compte l'irréversibilité et l'évolution de Systèmes Complexes Adaptatifs comme le souligne PRIGOGINE dans [200].

de la notion de perspectivisme [108], reflexions plus fine autour de la nature des modèles [252]². Les potentialités théoriques et empiriques de telles approches jouent nécessairement un rôle dans leur succès³, comme le confirme les domaines très variés d'application (voir [185] pour une revue très générale), comme par exemple la Science de Réseaux [16]; les Neurosciences [143]; les Sciences Sociales; la Géographie [175][202]; la Finance avec les approches éconophysiques [238]; l'Ecologie [117]. La Feuille de Route des Systèmes Complexes [43] propose une double lecture des travaux en Complexité : une approche horizontale faisant la connexion entre champs d'étude par des questions transversales sur les fondations théoriques de la complexité et des faits stylisés empiriques communs, et une approche verticale, dans le but de construire des disciplines intégrées et les modèles multi-scalaires hétérogènes correspondants. L'interdisciplinarité est ainsi cruciale pour notre contexte scientifique.

INTERDISCIPLINARITÉ

Il est important d'insister sur le rôle de l'interdisciplinarité dans la position de recherche prise ici. Il s'agit moins d'un travail en Géographie ou en Modélisation de Systèmes Complexes Adaptatifs, mais en *Science des Systèmes Complexes* qui se réclame disciplines propre comme le propose PAUL BOURGINE. Ce n'est pas sans risques d'être lu avec méfiance voir défiance par les tenants des disciplines classiques, comme des exemples récents de malentendus ou conflits ont récemment illustré [97]. Le positionnement de BATTY lorsqu'il propose *Une Nouvelle Science des Villes* [23] (qu'il présente avec humour comme *La nouvelle science des villes*), se présente comme une intégration des disciplines et méthodes vers une science définie par son objet d'étude, les villes.

L'évolution scientifique des sciences de la complexité, qui est vue par certains comme une révolution [66], ou même comme *un nouveau type de science*, pourrait affronter des difficultés intrinsèques dues aux comportements et a-priori des chercheurs en tant qu'être humains. Plus précisément, le besoin d'interdisciplinarité qui fait la force des Sciences de la Complexité pourrait devenir une de ses grandes fai-blesses, puisque la structure fortement en silo de la science peut avoir des impacts négatifs sur les initiatives impliquant des disciplines variées. Nous n'évoquons pas les problèmes de sur-publication, quantification, compétition, qui sont plus liés à des questions de Science Ouverte et de son éthique, tout aussi de grande importance mais d'une

² dans ce cadre, les progrès scientifiques et épistémologiques ne peuvent pas être dissociés et peuvent être vus comme étant en co-évolution

³ même si l'adoption de nouvelles pratiques scientifiques est souvent largement biaisé par l'imitation et le manque d'originalité [90], ou de façon plus ambivalente, par des stratégies de positionnement puisque le combat pour les fonds est un obstacle croissant à une recherche saine [35].

autre nature. Cette barrière qui nous hante et que nous pourrions ne pas surmonter, a pour plus évident symptôme des *divergences culturelles disciplinaires*, et les conflits d’opinion en résultant. Ce drame du malentendu scientifique est d’autant plus grave qu’il peut en effet détruire totalement certains progrès en interprétant comme une falsification des travaux qui traitent une question toute différente. L’exemple récent d’un travail sur les inégalités liées aux hauts revenus présenté dans [4], et dont les conclusions ont été commentées comme s’opposant aux thèses de Piketty dans [197], est typique de ce schéma. Alors que Piketty se concentre sur la construction de bases de données propres sur le temps long pour les revenus et montre empiriquement une récente accélération des inégalités de revenus, son modèle visant à lier ce fait stylisé avec l’accumulation de capital a été critiqué comme sur-simplifié. D’autre part, Bergeaud *et al.* montrent par un modèle d’économie de l’innovation que *sous certaines hypothèses* les écarts de revenus peuvent être bénéfique à l’innovation et donc à une utilité globale. D'où des conclusions divergentes sur le rôles des capitaux personnels dans une économie. Mais des *point de vus* ou *interprétations* différentes ne signifient pas une incompatibilité scientifique, et on pourrait même imaginer rassembler ces deux approches dans un cadre et modèle unifié, produisant des interprétations possiblement similaires et potentiellement encore nouvelles. Une telle approche intégrée aura de grandes chances de contenir plus d’information (selon comment le couplage est opéré) et être une avancée scientifique. Cette expérience de pensée illustre les potentialités et la nécessité de l’interdisciplinarité. Dans une autre veine assez similaire, [133] ré-analyse des données biologiques d’une expérience de 1943 qui prétendait confirmer l’hypothèse des processus d’évolution Darwiniens par rapport aux processus Lamarckiens, et montrent que les conclusions ne tiennent plus dans le contexte actuel d’analyse de données (avances énormes sur la théorie et les possibilités de traitement) et scientifique (avec d’autres nombreuses preuves de nos jours des processus Darwiniens) : c’est un bon exemple de malentendu sur le contexte, et comment le cadre de travail à la fois technique et thématique influence fortement les conclusions scientifiques. Nous développons à présent divers exemples révélateurs de la manière dont des conflits entre disciplines peuvent être dommageables.

LA PHYSIQUE RÉINVENTE LA GÉOGRAPHIE Comme déjà mentionné, DUPUY et BENGUIGUI soulignent dans [97] le fait que les sciences urbaines ont récemment connu des conflits ouverts entre les tenants classiques des disciplines et des nouveaux arrivants, en particulier les physiciens. La disponibilité de grand jeux de données d’un nouveau type (réseaux sociaux, données des nouvelles technologies de la communication) ont attiré leur attention sur des objets plus traditionnellement étudiés par les sciences humaines, puisque les méthodes

analytiques et computationnelles de la physique statistique sont devenues applicables. Bien que ces travaux soient généralement présentés comme la construction d'une approche scientifique des villes, tout en impliquant que la connaissance existante n'est pas scientifique de par sa nature plus qualitative, ils n'ont aucunement révélé de connaissance nouvelle sur les systèmes urbains : pour citer quelques exemples, [21] conclut que Paris a subit une transition pendant la période d'Haussman et ses opérations de planification globale, qui sont des faits naturellement connus depuis longtemps en Histoire Urbaine et Géographie Urbaine. [61] redécouvre que le modèle gravitaire est amélioré par l'introduction de décalages dans les interactions et dérive analytiquement l'expression d'une force d'interaction entre les villes, sans aucun cadre théorique ni thématique. De tels exemples peuvent être multipliés, confirmant l'inconfort courant entre physiciens et géographes. Des bénéfices significatifs pourraient résulter d'une intégration raisonnée des disciplines [188] mais la route semble être bien longue encore.

ECONOMIE GÉOGRAPHIE OU GÉOGRAPHIE ECONOMIQUE ? Des conflits similaires se rencontrent en économie : comme décrit par [178], la discipline de l'économie géographique, traditionnellement proche de la géographie, a fortement critiqué un nouveau courant de pensé nommé *économie géographisée*, dont le but est la spatialisation des techniques économiques classiques. Chacune n'ont pas les mêmes desseins et buts, et le conflit apparaît comme un malentendu complet vu d'un oeil extérieur.

MODÉLISATION BASÉE AGENT EN ECONOMIE Des conflits disciplinaires peuvent aussi se manifester sous la forme d'un rejet de méthodes nouvelles par les courants dominants. Suivant FARMER [100], l'échec opérationnel de la plupart des approches économiques classiques pourrait être compensé par un usage plus systématique de la modélisation et simulation basées agent. L'absence de cadre analytique qui est naturelle pour l'étude de la plupart des systèmes complexes adaptatifs semble rebuter la plupart des économistes.

FINANCE En finance quantitative coexistent divers champs de recherche ayant très peu d'interactions entre eux. On peut considérer deux exemples. D'une part, les statistiques et l'économétrie sont extrêmement avancées en mathématiques théoriques, utilisant par exemple des méthodes de calcul stochastique et de théorie des probabilités pour obtenir des estimateurs très raffinés de paramètres pour un modèle donné (voir par exemple [17]). D'autre part, l'éconophysiologie a pour but d'étudier des faits stylisés empiriques et inférer les lois correspondantes pour tenter d'expliquer les phénomènes liés à la complexité des marchés financiers [238], comme par exemple les

cascades menant aux ruptures de marché, les propriétés fractales des signaux des actifs, la structure complexe des réseaux de corrélation. Chacun a ses avantages dans un contexte particulier et gagnerait à des interactions accrues entre les deux domaines.

Ces divers exemples pris au fil du vent sont de brèves illustrations du caractère crucial de l'interdisciplinarité et de sa difficulté à pratiquer. Sans presque exagérer, on pourrait imaginer l'ensemble des chercheurs se plaindre de mauvaises ou difficiles expériences d'interdisciplinarité, avec un retour largement positif lors des rares succès. Nous allons tenter par la suite d'emprunter ce chemin étroit, empruntant des idées, théories et méthodes de diverse disciplines, dans l'idéal de la construction d'une connaissance intégrée. En effet, le couplage d'approches hétérogènes à différents niveaux et échelles sera une clé de voute de cette thèse, la moelle épinière de la philosophie sous-jacente et une composante de la théorie qu'on construira.

PARADIGMES DE LA COMPLEXITÉ EN GÉOGRAPHIE

Pour revenir à notre anecdote introductive, nous nous concentrerons sur l'étude d'un objet thématique qui sera les systèmes territoriaux. Plus généralement, il s'agit par commencer de brosser une revue du rôle de la complexité en géographie. Les géographes sont familiers avec la complexité depuis un certain temps, puisque l'étude des interactions spatiales est l'un de ses objets de prédilection. La variété de champs en géographie (géomorphologie, géographie physique, géographie environnementale, géographie humaine, géographie de la santé, etc. pour en nommer quelques) a sûrement joué un rôle clé dans la constitution d'une pensée géographique subtile, qui considère des processus hétérogènes et multi-scalaires.

PUMAIN rappelle dans [203] une histoire subjective de l'émergence des paradigmes de la complexité en géographie. La cybernétique a produit des théories des systèmes comme celle utilisée par Forrester. Plus tard, le glissement vers les concepts de criticalité auto-organisée et d'auto-organisation en physique ont conduit aux développements correspondants en géographie, comme [229] qui témoigne de l'application des concepts de la synergétique aux dynamiques des systèmes urbains. Enfin, les paradigmes actuels des systèmes complexes se sont introduits par plusieurs entrées. Par exemple, la nature fractale de la forme urbaine a été introduite par [24] et a eu de nombreuses applications jusqu'à des développements plus récents [141]. BATTY a aussi introduit les automates cellulaires en modélisation urbaine et propose une synthèse jointe avec les modèles basés agents et les fractales dans [22]. Une autre introduction de la complexité en géographie fut pour le cas des systèmes urbains à travers la théorie évolutive des villes de PUMAIN. En interaction intime avec la modélisation dès ses débuts (le premier modèle Simpop décrit par [230]

rentre dans le cadre théorique de [202]), cette théorie vise à comprendre les systèmes de villes comme des systèmes d'agents adaptatifs en co-évolution, aux interactions multiples, avec différents aspects mis en valeur comme l'importance de la diffusion des innovations. La série des modèles Simpop [206] a été conçue pour tester différentes hypothèses de la théorie. Par exemple, des processus sous-jacent différents ont été mis en évidence pour les systèmes de ville en Europe et aux Etats-unis [48]. A d'autres échelles de temps et dans d'autres contextes, le modèle SimpopLocal [233] a pour but d'étudier les conditions pour l'émergence de systèmes urbains hiérarchiques à partir d'établissements disparates. Un modèle minimal (au sens de paramètres nécessaires et suffisants) a été isolé grâce à l'utilisation de calcul intensif via le logiciel d'exploration de modèles OpenMole [234], ce qui était un résultat impossible à atteindre de manière analytique pour un tel type de modèle complexe. Les progrès techniques d'OpenMole [223] ont été menés simultanément avec les avances théoriques et empiriques. Les avancées épistémologiques ont également été cruciales dans ce cadre, comme REY le développe dans [224], et de nouveaux concepts comme la modélisation incrémentale [72] ont été découverts, avec de puissantes applications concrètes : [70] l'applique sur le système de villes soviétique et isole les processus socio-économiques dominants, par un test systématique des hypothèses thématiques et des fonctions d'implémentation. Des directions pour le développement de telles pratiques de Modélisation et Simulation en géographie quantitative ont récemment été introduits par BANOS dans [12]. Il conclut par neuf principes⁴, parmi lesquels on peut citer l'importance de l'exploration intensive des modèles computationnels et l'importance du couplage de modèles hétérogènes, qui sont avec d'autre principes tel la reproductibilité au centre de l'étude des systèmes complexes géographiques selon le point de vue décrit précédemment. Nous nous positionnons dans l'héritage de cette ligne de recherche, travaillant de manière conjointe sur les aspects théoriques, empiriques, épistémologiques et de modélisation.

QUESTION DE RECHERCHE

La question de recherche et les objets précis sont délibérément flous pour l'instant, puisque nous postulons que la construction d'une problématique ne peut être dissociée de la production d'une théorie correspondante. De manière réciproque, il n'y a aucun sens à poser des questions sorties de nulle part, sur des objets qui ont été seulement partiellement ou brièvement définis. Notre question préliminaire pour entrer dans le sujet, qu'on peut obtenir à partir de

⁴ Je me rappelle RENÉ DOURSAT insister pour la recherche du dernier commandement de BANOS

cas concrets comme l'anecdote introductory ou la revue de littérature préliminaire, est la suivante :

Comment définir les systèmes territoriaux, et les échelles et ontologies associées, dans une théorie cohérente, innovante et informative sur les processus sous-jacents ?

Il s'agit bien sûr d'une fausse question à ce stade, mais qui est toujours utile pour diriger la compréhension globale et le lecteur soucieux d'une démarche linéaire classique.

En effet, une caractéristique fondamentale des systèmes territoriaux est leur nature spatio-temporelle, qui est contenue dans leur dynamiques spatio-temporelles. La notion de *processus* au sens de [135] capture de plus les relations causales entre composantes de ces dynamiques, et est ainsi une approche intéressante pour une compréhension voire explication de ces systèmes. L'*échelle* doit être comprise ici au sens opérationnel (caractéristiques physiques) end l'*ontologie* comme les objets réels étudiés⁵. Notre question peut être vue grossièrement comme la recherche de théories et modèles qui révèlent des processus impliqués dans des systèmes complexes contenant aux moins des établissements humains, ce dernier point étant crucial pour la construction d'une problématique convergente plutôt que de se perdre dans des propositions irréalistes et non constructives qui pourrait aller de comprendre tout du cerveau (qui peut être vu comme une brique élémentaire des systèmes territoriaux qui émergent des interactions sociales) à l'écosphère qui inclut aussi les systèmes territoriaux.

CONTENU

This provisory Memoire is organized the following way. A first part with four chapters sets the thematic, theoretical and methodological background. The study of geographical systems implies, because of their complexity, a subtle combination of Theoretical constructions and Empirical Analysis, either in an inductive reasoning or in a didactic constitution of knowledge. The first part aims to approach our subject from the theoretical and methodological point of view, and rather as a *necessary foundation* shall be understood as a body of knowledge *coevolving* with Empirical and Modeling Parts. A linear reading is not necessarily the best way to deeply perceive the implications of theory

⁵ cet usage de la notion d'*ontologie* biaise naturellement la recherche vers des paradigmes de modélisation puisque qu'elle est proche de celle utilisée dans [161], mais nous prenons la position (développée en détails plus loin) de comprendre toute construction scientifique comme un *modèle*, rendant la frontière entre théories et modèles moins pertinentes que pour des visions plus classiques. Toute théorie doit faire des choix sur les objets décrits, leur relations et les processus impliqués, et contient donc une *ontologie* dans ce sens.

on empirical and modeling experiments and reciprocally. Some methodological developments are necessary but explicit reference will be done when it will be the case. A first chapter starts from the provisory research question given above and frames from a thematic point of view geographical objects and processes to be studied, resulting in precise research questions. The scene is set up for the construction of our theoretical background in a second chapter, that consists in a geographical theory for territorial systems on the one hand and in an epistemological theory of socio-technical systems modeling that frames our approach at a meta-level. We then develop methodological considerations on diverse questions implied by theory and required for modeling. Finally, a chapter of quantitative epistemology finishes to pave the way for modeling directions, unveiling literature gaps precisely linked to our question. A second part develops results obtained from empirical analysis and modeling experiments, along with on-going and planned projects in these fields. It first present empirical analysis aimed at identifying stylized facts. Toy-models of urban growth are then proposed, followed by an example and propositions for more complex models. The third part constructs our research objective for the remaining part of our project and sets a corresponding roadmap. Appendices contain non-digest important parts of our work such as models implementation architecture and details and specific tools developed for a reproducible research workflow.

Première partie

FOUNDATIONS

This part set up foundations, constructing our research precise subject and questions from a thematic point of view, completed with a theoretical construction for framing at thematic and epistemological levels. We also provide methodological digressions, and a quantitative epistemological analysis completing the manual state of the art.

INTERACTIONS BETWEEN NETWORKS AND TERRITORIES

Si la question de la priorité de l'œuf sur la poule ou de la poule sur l'œuf vous embarrasse, c'est que vous supposez que les animaux ont été originaiement ce qu'ils sont à présent.

- DENIS DIDEROT [89]

Cette analogie est idéale pour introduire les notions de causalité et de processus dans les systèmes territoriaux. En voulant traiter naïvement des questions similaires à notre question de recherche préliminaire, certains qualifiés les causalités au sein de systèmes complexes comme un problème “de poule et œuf” : si un effet semble causer l'autre et réciproquement, comment est-il possible d'isoler les processus correspondants ? Cette vision est souvent présente dans les approches réductionnistes qui ne postulent pas une complexité intrinsèque au sein des systèmes étudiés. L'idée suggérée par DIDEROT est celle de *co-evolution* qui est un phénomène central dans les dynamiques évolutionnaires des Systèmes Complexes Adaptatifs comme HOLLAND élabore dans [132]. Il fait le lien entre la notion d'émergence (ignorée dans les approches réductionnistes), en particulier l'émergence de structures à une plus grande échelle par les interactions entre agents à une échelle donnée, en général concrétisée par un système de limites, qui devient cruciale pour la co-évolution des agents à toutes les échelles : l'émergence d'une structure sera simultanée avec une autre, chacune exploitant leur interrelations et environnements générés conditionnés par le système de limites. Nous explorerons ces idées pour le cas des systèmes territoriaux par la suite.

Ce chapitre introductif est destiné à poser le cadre thématique, le contexte géographique sur lesquels les développements suivants se baseront. Il n'est pas supposé être compris comme une revue de littérature exhaustive ni comme les fondations théoriques fondamentales de notre travail (le premier point étant l'objet du chapitre 3 tandis que le second sera traité plus tôt dans le chapitre 8), mais plutôt comme une construction narrative ayant pour but d'introduire nos objets et positions d'étude, afin de construire naturellement des questions de recherche précises.

1.1 RÉSEAUX ET TERRITOIRES

1.1.1 *Une circularité naturelle*

TERRITORIALITÉ HUMAINE Une entrée possible dans l'ensemble des objets géographiques que nous proposons d'étudier est la notion de territoire. En Ecologie, un territoire correspond à l'étendue spatiale occupée par un groupe d'agent ou plus généralement un écosystème. Les *Territoires Humains* sont extrêmement plus complexes de par l'importance de leur représentations sémiotiques, qui jouent un rôle significatif dans l'émergence des constructions sociétales. Selon RAFFESTIN dans [214], la *Territorialité Humaine* est "la conjonction d'un processus territorial avec un processus informationnel", ce qui implique que l'occupation physique et l'exploitation de l'espace par les sociétés humaines n'est pas dissociable des représentations (cognitives et matérielles) de ces processus territoriaux, qui influent en retour leur évolution. En d'autres termes, à partir de l'instant où les constructions sociales déterminent la constitution des établissements humains, les structures sociales abstraites et concrètes joueront un rôle dans l'évolution des systèmes territoriaux, par exemple à travers la propagation d'informations et de représentations, par des processus politiques, ou encore par la correspondance effective entre territoire vécu et territoire perçu. Bien que cette approche ne donne pas de conditions explicites pour l'émergence d'un système séminal d'établissements agrégés (c'est à dire l'émergence des villes), elle insiste sur leur rôle comme lieu de pouvoir et de création de richesse au travers des échanges. Mais la ville n'a pas d'existence sans son hinterland et le système territorial peut difficilement être résumé par ses villes, comme un système de villes. En se restreignant à ce sous-système, il y a toutefois compatibilité entre la théorie de territoires de RAFFESTIN et la théorie évolutive des villes de PUMAIN [205], qui interprète les villes comme des systèmes complexes dynamiques auto-organisés, qui agissent comme des médiateurs du changement social : par exemple, les cycles d'innovation s'initialisent au sein des villes et se propagent entre elles. Les villes sont ainsi des agents compétitifs qui co-évoluent (au sens donné précédemment). Le système territorial peut ainsi être compris comme une structure sociale organisée dans l'espace, qui comprend ses artefacts concrets et abstraits. Une étendue spatiale imaginaire avec des ressources potentielles qui n'aurait jamais connu de contact avec l'humain ne pourra pas être un territoire si elle n'est pas habitée, imaginée, vécue, exploitée, même si ces ressources pourraient être potentiellement exploitée le cas échéant. En effet, ce qui est considéré comme une ressource (naturelle ou artificielle) dépendra de la société (par exemple de ses pratiques et de ses capacités technologiques). Un aspect central des établissements humains qui a une longue tradition d'étude en géographie, et qui est

directement relié à la notion de territoire, est celui des *réseaux*. Nous allons voir comment le passage de l'un à l'autre est inévitable et leur définition indissociable.

UNE THÉORIE TERRITORIALE DES RÉSEAUX Nous paraphrasons DUPUY dans [96] lorsqu'il propose des éléments pour une "théorie territoriale des réseaux" basée sur le cas concret d'un réseau de transport urbain. Cette théorie présente les *réseaux réels* (i.e. les réseaux concrets, incluant les réseaux de transport) comme la matérialisation de *réseaux virtuels*. Plus précisément, un territoire est caractérisé par de fortes discontinuités spatio-temporelles induites par la distribution non-uniforme des agents et des ressources. Ces discontinuités induisent naturellement un réseau de "projets transactionnels" qui peuvent être compris comme des interactions potentielles entre les éléments du système territorial (agents et/ou ressources). Par exemple, de nos jours les actifs se doivent d'accéder à la ressource qu'est l'emploi, et des échanges économiques s'effectuent entre les différents territoires spécialisés dans les productions de différents types. En tout temps des interactions potentielles ont existé¹ Le réseau d'interaction potentiel est concrétisé quand l'offre s'adapte à la demande, et résulte en la combinaison de contraintes économiques et géographiques avec les motifs de demande, de manière non-linéaire via des agents qu'on peut désigner comme *opérateurs*. Un tel processus est loin d'être immédiat, et conduit à de forts effets de non-stationnarité et de dépendance au chemin : l'extension d'un réseau existant dépendra de la configuration précédente, et selon les échelles de temps impliquées, la logique et même la nature des opérateurs peut avoir évolué. RAFFESTIN souligne dans sa préface de [191] qu'une théorie géographique articulant espaces, réseaux et territoires n'a jamais été formulée de manière cohérente. Il semble que c'est toujours le cas aujourd'hui, même si la théorie évoquée ci-dessus semble être un bon candidat bien qu'elle reste à un niveau conceptuel. La présence d'un territoire humain implique nécessairement la présence de réseaux d'interactions abstraites et de réseaux concrets utilisés pour transporter les individus et les ressources (incluant les réseaux de communication puisque l'information est une ressource essentielle). Selon le régime dans lequel le système considéré se trouve, le rôle respectif du réseau peut être radicalement différent. Selon DURANTON [98], les villes pré-industrielles étaient limitées en croissance de par les limitations des réseaux de transport. Les progrès technologiques ont permis de les surmonter et à mené à la prépondérance du marché foncier dans la formation des villes (et par conséquent un rôle des réseaux de transport qui déterminent les prix par l'accessibilité), et

¹ même quand le nomadisme devait encore être la règle, des réseaux d'interactions potentielles dynamiques dans l'espace ont du exister, mais devaient avoir moins de chance de se matérialiser en des routes matérielles.

plus récemment à une importance croissante des réseaux de télécommunication ce qui a induit une "tyrannie de la proximité" puisque la présence physique n'est pas remplacable par une communication virtuelle. Cette approche territoriale des réseaux semble naturelle en géographie, puisque les réseaux sont étudiés conjointement avec des objets géographiques auxquels est associée une théorie, en opposition à la science des réseaux qui étudie brutalement les réseaux spatiaux avec peu de fond thématique [95].

DES RÉSEAUX QUI FAÇONNENT LES TERRITOIRES ? Cependant les réseaux ne sont pas seulement une manifestation matérielle de processus territoriaux, mais jouent également leur rôle dans ces processus comme leur évolution peut influencer l'évolution des territoires en retour. Dans le cas des *réseaux techniques*, une autre désignation des réseaux réels donnée dans [191], de nombreux exemples de tels retroactions peuvent être mis en évidence : l'interconnexion des réseaux de transport permet des motifs de mobilité multi-échelles, formant ainsi le territoire vécu. A une plus petite échelle, des changements de l'accessibilité peuvent induire l'adaptation d'un espace fonctionnel urbain. Il émerge alors une difficulté intrinsèque : il est loin d'évident d'attribuer des mutations territoriales à une évolution du réseau and réciproquement la matérialisation d'un réseau à des dynamiques territoriales précises. Revenir à la citation de Diderot devrait aider à ce point, au sens où il ne faut pas considérer le réseau ni les territoires comme des systèmes indépendants qui s'influencerait mutuellement par des relations causales, mais comme des composantes fortement couplées d'un système plus large. La confusion autour de possibles relations causales simples a nourri un débat scientifique encore actif aujourd'hui. Les méthodologies pour identifier ce qui est nommé *effets structurants* des réseaux de transport ont été proposées par les planificateurs dans les années 1970 [39, 40]. Il aura fallu un certain temps pour un positionnement critique sur l'usage non raisonné et decontextualisé de ces méthodes par les planificateurs et les politiques généralement pour justifier technocratiquement des projets de transports. Cela a été fait en premier par OFFNER dans [190]. Récemment un édition spéciale du même journal sur ce débat [148] a rappelé d'une part que les mauvaises interprétations et les mauvais usages étaient encore largement présent aujourd'hui dans les milieux opérationnels de la planification comme [78] confirme, et d'autre part qu'il faudrait encore une certaine quantité de progrès scientifique pour comprendre en profondeur les relations entre réseaux et territoires. PUMAIN souligne que des travaux récents ont révélé des effets systémiques sur de très longues échelles temporelles (comme e.g. le travail de BRETAGNOLLE sur l'évolution des chemins de fer, qui montre une sorte d'effet structurel sur la nécessité de connexion au réseau des villes, afin de rester actives, mais qui n'est ni suffisant ni totalement

causal). A un niveau macroscopique des motifs typiques d'interaction émergent, mais les trajectoires microscopiques du systèmes sont essentiellement chaotiques : la compréhension des dynamiques couplées dépend fortement de l'échelle considérée. A une petite échelle il est peu raisonnable de vouloir montrer des comportement systématisques, comme le rappelle OFFNER. Par exemple, sur des territoires de montagne français comparables, [31] montre que les réactions à un même contexte d'évolution du réseau de transport peut mener à des réactions territoriales très diverses, certains trouvant de forts bénéfices par la nouvelle connectivité, d'autres au contraire devenant plus fermés. Ces retroactions potentielles des réseaux sur les territoires n'agit pas nécessairement sur des composantes concrètes : CLAVAL montre dans [65] que les réseaux de transport et de communication contribuent à la représentation collective d'un territoire en agissant sur un sentiment d'appartenance.

SYSTÈMES TERRITORIAUX Ce voyage des territoires aux réseaux, et retour, nous permet d'esquisser une définition préliminaire d'un système territorial sur laquelle se basera les considérations théoriques suivantes. Comme nous avons mis en exergue le rôle des réseaux, la définition se doit de les prendre en compte.

Définition provisoire. *Un Système Territorial est un territoire humain auquel peuvent être associés à la fois un réseau d'interactions et un réseau réel. Les réseaux réels sont une composante à part entière du système, jouant dans les processus d'évolution, au travers de multiples retroactions avec les autres composantes à plusieurs échelles spatiales et temporelles.*

Cette lecture des systèmes territoriaux est conditionnée à l'existence des réseaux et pourrait écarter certains territoires humains, mais il s'agit d'un choix délibéré justifié par les considérations précédentes, et qui précise notre sujet vers l'étude des interactions entre réseaux et territoires.

1.1.2 Réseaux de Transport

LA PARTICULARITÉ DES RÉSEAUX DE TRANSPORT Déjà évoqués dans le cas des effets structurants des réseaux, les réseaux de transports jouent un rôle déterminant dans l'évolution des territoires. Même si d'autres types de réseaux sont également fortement impliqués dans l'évolution des systèmes territoriaux (voir e.g. les débats sur l'impact des réseaux de communication sur la localisation des activités économiques), les réseaux de transport conditionnent d'autres types de réseaux (logistique, échanges commerciaux, interactions sociales concrètes pour donner quelques exemples) and semblent dominer dans les motifs d'évolution territoriale, en particulier dans nos sociétés contemporaines qui sont devenues dépendantes des réseaux

de transport [25]. Le développement du réseau français à grande vitesse est une illustration pertinente de l'impact des réseaux de transport sur les politiques de développement territorial. Présenté comme une nouvelle ère de transport sur rail, une planification par le haut de lignes totalement nouvelles a été présenté comme central pour le développement [273]. Le manque d'intégration de ces nouveaux réseaux avec l'existant et avec les territoires locaux est à présent observé comme une faiblesse structurelle et des impacts négatifs sur certains territoires ont été prouvés [274]. Une revue faite dans [26] confirme qu'aucune conclusion générale sur des effets locaux d'une connection à une ligne à grande vitesse ne peut être tirée, bien que ce sésame garde une place conséquente dans les imaginaires des élus. Ces exemples illustrent comment les réseaux de transport peuvent avoir des effets à la fois directs et indirects sur les dynamiques territoriales. La planification intégrée, au sens d'une planification coordonnée entre les infrastructures de transport et le développement urbain, considère le réseau comme une composante déterminante du système territorial. Les Villes Nouvelles parisiennes sont un tel cas qui témoigne de la complexité de ces actions de planification qui le plus souvent ne mène pas au effets initialement désirés [193]. Des projets récents comme [149] ont tenté d'implémenter des idées similaires, mais il manque pour l'instant de recul pour juger de leur succès à produire un territoire effectivement intégré. Les réseaux de transports sont dans tous les cas au centre de ces approches des territoires urbains. Nous nous concentrerons par la suite sur les réseaux de transport pour toutes ces raisons évoquées ici.

DÉCONSTRUIRE L'ACCESSIBILITÉ La notion d'accessibilité surgit rapidement lorsqu'on s'intéresse aux réseaux de transport. Basée sur la possibilité d'accéder un lieu par un réseau de transport (pouvant prendre en compte la vitesse, la difficulté de se déplacer), elle est généralement définie comme un potentiel d'interaction spatiale² [25]. Cet objet est souvent utilisé comme un outil de planification ou comme une variable explicative de localisation des agents par exemple. Il faut cependant rester prudent sur son usage inconditionnel. Plus précisément, il peut s'agir d'une construction qui ignore une partie conséquente des dynamiques territoriales. La mystification de la notion de *mobilité* a été montrée par COMMENGES dans [68], qui révèle que la majorité des débats sur la modélisation de la mobilité et les notions correspondantes était majoritairement construites de manière ad-hoc par les administrateurs de transports issus du *Corps des Ponts* qui importaient brutalement les outils et méthodes des Etats-Unis sans adaptation ni reflexion adaptée au contexte français. L'accessibilité

² et souvent généralisée comme une *accessibilité fonctionnelle*, par exemple les emplois accessibles aux actifs d'un lieu. Les potentiels d'interaction spatiaux s'exprimant dans les lois de gravité peuvent aussi être compris de cette façon.

pourrait de même être une construction sociale et n'avoir que peu de fondement théorique, puisqu'il s'agit en grande partie d'un outil de modélisation et de planning. Les débats récents sur la planification du *Grand Paris Express* [173], cette nouvelle infrastructure de transport métropolitaine

ECHELLES ET HIERARCHIES

INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES At this state of progress, we have naturally identified a research subject that seems to take a significant place in the complexity of territorial systems, that is the study of interactions between transportation networks and territories. In the frame of our preliminary definition of a territorial system, this question can be reformulated as the study of networked territorial systems with an emphasize on the role of transportation networks in system evolution processes.

1.2 MODÉLISER LES INTERACTIONS

1.2.1 Modélisation en Géographie Quantitative

La modélisation en Géographie Théorique et Quantitative (TQG), et plus généralement en Sciences Sociales, a une longue histoire dont nous ne pourrons que brosser un bref portrait ici. CUYALA procède dans [80] à une analyse spatio-temporelle du mouvement de la Géographie Théorique et Quantitative en langue française et souligne l'émergence de la discipline comme une combinaison d'analyses quantitatives (e.g. analyse spatiale et pratiques de modélisation et de simulation) et de construction théoriques. L'intégration de ces deux composantes permet la construction de théories à partir de faits stylisés empiriques, qui produisent à leur tour des hypothèses théoriques pouvant être testées sur les données empiriques. Cette approche est née sous l'influence de la *New Geography* dans les pays Anglo-saxons et en Suède. Une histoire étendue de la genèse des modèles de simulation en géographie est faite par REY dans [224] avec une attention particulière pour la notion de validation de modèles. L'utilisation de ressources de calcul pour la simulation de modèles est antérieur à l'introduction des paradigmes de la complexité, remontant à HÄGERSTRAND et FORRESTER, pionniers des modèles d'économie spatiale inspirés par la cybernétique. Avec l'augmentation des potentialités de calcul, des transformations épistémologiques ont également suivi, avec l'apparition de modèles explicatifs comme outils expérimentaux. REY compare le dynamisme des années soixante-dix quand les centres de calcul furent ouverts aux géographes à la démocratisation actuelle du Calcul Haute Performance (calcul sur grille à l'utilisation transparente, voir [234] pour un exemple des possibilités offertes en terme de calibration et de validation de modèle, réduisant le temps de calcul nécessaire de 30 ans à une semaine), qui est également accompagnée par une évolution des pratiques [12] et techniques [63] de modélisation. La modélisation, et en particulier les modèles de simulation, est vue par beaucoup comme une brique fondamentale de la connaissance : [161] rappelle la combinaison des domaines empirique, conceptuel (théorique) et de la modélisation, avec des retroactions constructives entre chaque. Une modèle peut être un outil d'exploration pour tester des hypothèses, un outil empirique pour valider une théorie sur des jeux de données, un outil explicatif pour révéler des causalités et ainsi des processus internes au système, un outil constructif pour construire itérativement une théorie conjointement avec celle des modèles associés. Ce sont des exemples de fonctions parmi d'autres : Varenne donne dans [251] une classification raffinée des diverses fonctions d'un modèle. Nous considérons la modélisation comme un instrument fondamental de connaissance des processus au sein de systèmes complexes adaptatifs, et précisons encore

notre question de recherche, qui s'intéressera aux *modèles impliquant des interactions réseaux et territoires*.

1.2.2 Modéliser les territoires et réseaux

Au sujet de notre question précise des interactions entre réseaux de transport et territoires, nous proposons un aperçu des différentes approches. Selon [47], “les idées des spécialistes de la planification cherchant à donner des définitions des systèmes de ville, depuis 1830, sont étroitement liées aux transformations des réseaux de communication”. C'est en quelque sorte la prophétie auto-réalisatrice inversée, au sens où elle est déjà réalisée avant d'être formulée. Cela implique que les ontologies et les modèles correspondants proposés par les géographes et les planificateurs sont fortement liés aux préoccupations historiques courantes, ainsi forcément limités en portée et raisons. Dans une vision perspectiviste de la science [108] de telles limites sont l'essence de l'entreprise scientifique, et comme nous démontrerons en chapitre 8 leur combinaison et couplage dans le cas de modèles est une source de connaissance.

Modèles LUTI

Une partie importante de la littérature proposant des modélisations des interactions entre réseaux et territoires se trouve dans le domaine de la planification urbaine, avec les *modèles d'interaction entre usage du sol et transport (LUTI)*. Ces travaux peuvent être difficiles à cerner car liés à différentes disciplines. Par exemple, du point de vue de l'Economie Urbaine, les propositions de modèle intégrés existent depuis un certain temps [210]. La variété des modèles existants a conduit à des comparaisons opérationnelles [194, 260]. Plus récemment, les avantages respectifs des approches statiques et dynamiques a été étudié par [145]. Dans tous les cas, ce type de modèle opère généralement à des échelles temporelles et spatiales relativement faibles. [259] donne un état de l'art des études empiriques et de modélisation sur ce type d'approche des interactions entre usage du sol et transport. Le positionnement théorique est plutôt proche des disciplines de l'Economie, de la Planification et de la Sociologie, et relativement de nos raisonnements géographiques qui se veulent de comprendre également des processus sur le temps long. Pas moins de dix-sept modèles sont comparés et classifiés, parmi lesquels aucun n'inclut une évolution endogène du réseau de transport sur les échelles de temps relativement petites des simulations. Une revue complémentaire est faite par [57], élargissant le contexte avec l'inclusion de classes plus générales de modèles, comme des modèles d'interactions spatiales (parmi lesquels l'attribution du traffic et les modèles à quatre temps), les modèles de planification basés sur la recherche opérationnelle (optimisation des localisations), les modèles microscopiques d'utilité aléatoire, et

les modèles de marché foncier. Toutes ces techniques opèrent également à une petite échelle et considèrent au plus l'évolution de l'usage du sol. [136] couvre un horizon similaire avec une emphase supplémentaire sur les modèles à automates cellulaires d'évolution d'usage du sol et les modèles basés agent. Les modèles LUTI sont toujours largement étudiés et appliqués, comme par exemple [86] qui est utilisé pour la région métropolitaine parisienne. La courte portée temporelle d'application de ces modèles et leur nature opérationnelle les rend utiles pour la planification, ce qui est assez loin de notre souci d'obtenir des modèles explicatifs de processus géographiques.

Croissance du Réseau

La croissance de réseaux est pratiquée dans des entreprises de modélisation qui cherchent à expliquer de manière endogène la croissance des réseaux de transport, généralement d'un point de vue *bottom-up*, i.e. en mettant en évidence des règles locales qui permettraient de reproduire la croissance du réseau sur de longues échelles de temps (souvent le réseau de rues). Les économistes ont proposés des modèles de ce type : [276] passe en revue la littérature en économie de transports sur la croissance des réseaux dans le contexte d'une théorie endogène de la croissance [3], rappelant les trois aspects principalement traités par les économistes sur le sujet, qui sont la tarification routière, l'investissement en infrastructures et le régime de propriété, et propose finalement un modèle analytique combinant les trois. [267] propose une revue étendue de la modélisation de croissance des réseaux, en prenant en compte d'autres champs : la géographie des transports a développé très tôt des modèles basés sur des faits empiriques mais qui se sont concentrés sur reproduire la topologie plutôt que sur les mécanismes selon [267] ; les modèles statistiques sur des cas d'étude fournissent des conclusions très mitigées sur les relations causales entre offre et demande ; les économistes ont étudié la production d'infrastructure à la fois d'un point de vue microscopique et macroscopique, généralement non spatiaux ; la science des réseaux a produit des modèles jouet de croissance de réseau qui se basent sur des règles topologiques et structurelles plutôt que des règles se reposant sur des processus inspirés de faits réels. Une autre approche qui n'est pas mentionnée et que nous allons approfondir est la conception de réseau inspirée de la biologie. Nous donnons pour commencer des exemples d'études utilisant des concepts économiques ou géométriques pour modéliser la croissance de réseau. [272] montre avec un modèle économique basé sur des processus auto-renforçants et incluant une règle d'investissement basée sur l'attribution du trafic, que des règles locales sont suffisantes pour faire émerger une hiérarchie du réseau routier à usage du sol fixé. Une modèle très similaire donnée par [167] avec des fonctions coûts-bénéfices plus simples obtient une conclusion similaire. Alors que ces modèles basés sur des proce-

sus cherchent à reproduire des motifs macroscopiques des réseaux (typiquement les lois d'échelle), les modèles d'optimisation géométrique cherchent à ressembler à des réseaux réels dans leur topologie. [19] décrit un modèle basé sur une optimisation locale de l'énergie, mais ce modèle reste très abstrait et non validé. Le modèle de morphogenèse de [75] qui utilise des potentiels locaux et des règles de connectivité, même s'il n'est pas calibré, semble reproduire de manière plus raisonnable des motifs réels des réseaux de rues. Un modèle très proche est décrit dans [228]. D'autres tentatives comme [81, 270] sont plus proches de la modélisation procédurale [154, 258] et pour cette raison n'ont pas d'intérêt pour notre cas puisqu'ils peuvent difficilement être utilisés comme modèles explicatifs.

Modélisation Hybride

Les modèles de simulation qui incluent un couplage des dynamiques de la croissance urbaine et du réseau de transport sont relativement rares, et généralement pauvres d'un point de vue théorique et thématique. Une généralisation du modèle d'optimisation locale géométrique décrit précédemment a été développé dans [20]. Comme pour le modèle de croissance de réseau routier dont il est l'extension, les mécanismes locaux n'ont pas de justification théorique ou thématique, et le modèle n'est de plus pas exploré et aucune connaissance géographique ne peut en être tirée. [156] prend une approche économique plus intéressante, similaire à un modèle à quatre étapes (génération de flux origine-destination basés sur la gravité, attribution du trafic par Equilibre Utilisateur Stochastique) qui inclut coût de transport et congestion, couplé avec un module d'investissement routier qui simule les revenus des péages pour les agents qui construisent, et un module d'évolution d'usage du sol qui met à jour les actifs et emplois par modélisation de choix discrets. Les expériences montrent que l'usage du sol et le réseau en co-évolution mène à des retroactions positives renforçant les hiérarchies, mais sont loin d'être satisfaisantes pour deux raisons : d'une part la topologie du réseau n'évolue pas à proprement parler puisque seules les capacités et les flux changent dans le réseau, ce qui signifie que des mécanismes plus complexes sur de plus longues échelles de temps ne sont pas pris en compte, et d'autre part les conclusions sont assez limitées puisque le comportement du modèle n'est pas connu, les analyses de sensibilité étant faites sur un petit nombre d'espaces unidimensionnels : les mécanismes exhaustifs restent ainsi inconnus comme seuls des cas particuliers sont donnés dans l'analyse de sensibilité. D'un autre point de vue, [158] est aussi présenté comme un modèle de co-évolution mais correspond plus à une analyse statistique couplée puisqu'elle repose sur un modèle prédictif à chaîne de Markov. [227] décrit un modèle dans lequel le couplage entre usage du sol et la topologie du réseau est fait par un paradigme faible, l'usage du sol et l'accessibi-

lité n'ayant pas de retroaction sur la topologie du réseau. [2] décrit un modèle de co-évolution à une très petite échelle (échelle du bâtiment), dans lequel l'évolution du réseau et des bâtiments sont tous les deux régis par un agent commun (qui est influencé différemment par la topologie du réseau et la densité de population) ce qui implique une simplification trop grande des processus sous-jacents. Enfin, un modèle hybride simple exploré et appliqué à un exemple jouet de planification dans [217], repose sur les mécanismes d'accès aux activités urbaines pour la croissance des établissements avec un réseau s'adaptant à la forme urbaine. Les règles pour la croissance du réseau sont trop simples pour capturer les processus qui nous intéressent, mais le modèle produit à une petite échelle une large gamme de formes urbaines qui reproduisent les motifs typiques des établissements humains.

Modélisation de Systèmes Urbains

Une approche assez proche de nos questionnements courants est celle de la modélisation intégrée des systèmes de villes. Dans la continuité des modèles Simpop pour modéliser les systèmes de villes, SCHMITT décrit dans [233] le modèle SimpopNet qui vise à précisément intégrer les processus de co-évolution dans les systèmes de villes à longue échelle temporelle, typiquement par des règles pour un développement hiérarchique du réseau comme fonction des dynamiques des villes, couplées à celles-ci qui dépendent de la topologie du réseau. Malheureusement le modèle n'a pas été exploré ni étudié de manière plus approfondie, et de plus est resté au niveau de modèle jouet. COTINEAU propose une croissance endogène des réseaux de transport comme la dernière brique de construction de ses productions Marius mais cela reste à un niveau conceptuel. Nous nous positionnerons particulièrement dans cette lignée de recherche dans cette thèse.

1.2.3 Ebauche d'une Modélographie

An ongoing work is the production of a synthesis of this overview, from a modular modeling point of view, combined with a purpose and scale classification. Already mentioned, modular modeling consists in the integration of heterogeneous processes and implementation of processes in order to extract the set of mechanisms giving the best fit to empirical data [72]. We can thus classify models described here according to their building bricks in terms of processes implemented and thus identify possible coupling potentialities. This work is a preliminary step for the analysis in quantitative epistemology developed in chapter 3.

1.3 DE LA RECHERCHE QUALITATIVE : UNE EXPERIENCE EN OBSERVATION FLOTTANTE

Si le diable est dans les détails, les systèmes de transport entre autres sont l'allégorie de cette adage. Ce que certains appellent détail contient la majorité de l'information pour d'autres. Logiquement enfermés dans une bulle scientifique, malgré toutes les volontés développées en introduction, on tâchera de rester conscient de la nature et la portée de la connaissance produite ici. Ce que nous pourrions appeler détail, lors de l'étude de l'accessibilité d'un réseau de transport par exemple, tel des impressions ressenties par les usagers ou les relations sociales induites par les situations découlant des dynamiques du systèmes, seront le centre du questionnement pour un anthropologue ou sociologue. Une telle connaissance, qui trouverait certainement une place dans nos problématiques, est hors de notre portée de par l'absence de *terrain* de longue durée.

1.4 QUESTION DE RECHERCHE

To close this thematic touring introducing chapter, we can state a general research question that frames our further theoretical constructions and first modeling attempts. It is roughly the same as the problematic given at the end of previous section, but adding the insight of modeling as the approach to understand these complex systems.

General research Question. *To what extent a modeling approach to territorial systems as networked human territories can help disentangling complexly involved processes ?*

This question will be refined by theoretical developments in the next chapter and experiments in the followings.

2

METHODOLOGICAL DEVELOPMENTS

We are now building a rigorous Science of Cities, contrarily to what was done before.

- MARC BARTHÉLÉMY

Such a shocking phrase was pronounced during the introduction of a *Network* course for students of Complex System Science. Besides the fact that the spirit of CSS is precisely the opposite, i. e. the construction of integrative disciplines (vertical integration that is necessarily founded on the existing body of knowledge of concerned fields) that answer transversal questions (horizontal integration that imply interdisciplinarity) - see e. g. the roadmap for CS [43], it reveals how methodological considerations shape the perceptions of disciplines. From a background in Physics, "rigorous" implies the use of tools and methods judged more rigorous (analytical derivations, large datasets statistics, etc.). But what is rigorous for someone will not be for an other discipline¹, depending on the purpose of each piece of research (perspectivism [108] poses the *model*, that includes methods, as the articulating core of research enterprises). Thus the full role of methodology aside and not beside theory and experiments. We go in this chapter into various methodological developments which may be precisely used later or contribute to the global background.

We first propose a kind of essay insisting on the importance of reproducibility in science. More than a guideline, it is a way to practice science that a necessary condition for its rigor. Any non-reproducible work is not scientific. We then derive technical results on models of urban growth and on the sensitivity of scaling laws, that are both recurrent themes in the modeling of complex urban systems. We then introduce a method in the context of systematic model exploration and model behavior. We finally work on a link between static and dynamic correlations in a geographical system. This chapter is rather heterocline as sections may correspond to a particular technical need at a point in the thesis, to global methodological directions, or global research directions.

¹ a funny but sad anecdote told by a friend comes to mind : defending his PhD in statistics, he was told at the end by economists how they were impressed by the mathematical rigor of his work, whereas a mathematician judged that "he could have done everything on the back of an enveloppe".

2.1 REPRODUCIBILITÉ

The strength of science comes from the cumulative and collective nature of research, as progresses are made as Newton said “standing on the shoulder of giants”, meaning that the scientific enterprise at a given time relies on all the work done before and that advances would not be possible without constructing on it. It includes development of new theories, but also extension, testing or falsifiability of previous ones. In that context

As scientific reproducibility is an essential requirement for any study, its practice seems to be increasing [240] and technical means to achieve it are always more developed (as e.g. ways to make data openly available, or to be transparent on the research process such as git [220], or to integrate document creation and data analysis such as knitr [268]), at least in the field of numerical modeling and simulation. However, the devil is indeed in the details and obstacles judged at first sight as minor become rapidly a burden for reproducing and using results obtained in some previous researches. We describe two cases studies where models of simulation are apparently highly reproducible but unveil as puzzles on which research-time balance is significantly under zero, in the sense that trying to exploit their results may cost more time than developing from scratch similar models.

2.1.1 *Sur la Besoin d'expliquer le modèle*

A current myth is that providing entire source code and data will be a sufficient condition for reproducibility. It will work if the objective is to produce exactly same plots or statistical analysis, assuming that code provided is the one which was indeed used to produce the given results. It is however not the nature of reproducibility. First, results must be as much implementation-independent as possible for clear robustness purposes. Then, in relation with the precedent point, one of the purposes of reproducibility is the reuse of methods or results as basis or modules for further research (what includes implementation in another language or adaptation of the method), in the sense that reproducibility is not replicability as it must be adaptable [94].

Our first case study fits exactly that scheme, as it was undoubtedly aimed to be shared with and used by the community since it is a model of simulation provided with the Agent-Based simulation platform NetLogo [264]. The model is also available online [81] and is presented as a tool to simulate socio-economic dynamics of low-income residents in a city based on a synthetic urban environment, generated to be close in stylized facts from the real town of Tijuana, Mexico. Beside providing the source code, the model appears to be poorly documented in the literature or in comments and description of the

implementation. Comments made thereafter are based on the study of the urban morphogenesis part of the model (setup for the “residential dynamics” component) as it is our global context of study [216]. In the frame of that study, source code was modified and commented, which last version is available on the repository of the project².

FORMALISATION RIGOUREUSE An obvious part of model construction is its rigorous formalization in a formal framework distinct from source code. There is of course no universal language to formulate it [12], and many possibilities are offered by various fields (e.g. UML, DEVS, pure mathematical formulation). No paper nor documentation is provided with the model, apart from the embedded NetLogo documentation since it only thematically describes in natural language the ideas behind each step without developing more and provides information about role of different elements of the interface.

This formulation is a key for it to be understood, reproduced and adapted ; but it also avoids implementation biases such as

- Architecturally dangerous elements : in the model, world context is a torus and agents may “jump” in the euclidian representation, what is not acceptable for a 2D projection of real world. To avoid that, many tricky tests and functions were used, including unadvised practices (e.g. dead of agents based on position to avoid them jumping).
- Lack of internal consistence : the example of the patch variable `land-value` used to represent different geographical quantities at different steps of the model (morphogenesis and residential dynamics), what becomes an internal inconsistence when both steps are coupled when option `city-growth?` is activated.
- Coding errors : in an untyped language such as NetLogo, mixing types may conduct to unexpected runtime errors, what is the case of the patch variable `transport` in the model (although no error occurs in most of run configurations from the interface, what is more dangerous as the developer thinks implementation is secure). Such problems should be avoided if implementation is done from an exact formal description of the model.

IMPLÉMENTATION TRANSPARENTE A totally transparent implementation is expected, including ergonomics in architecture and coding, but

COMPORTEMENT ATTENDU DU MODÈLE Whatever the definition, a model can not be reduced to its formulation and/or implementation, as expected model behavior or model usage can be viewed

² at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite>

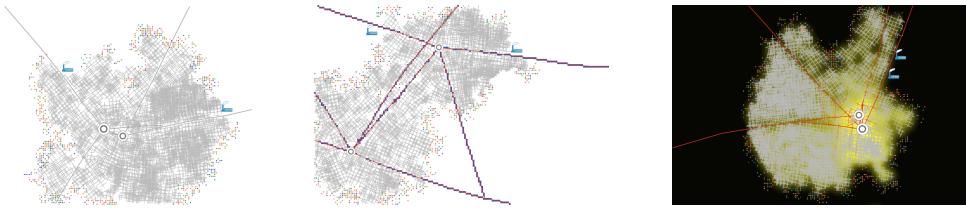


FIGURE 1 :

as being part of the model itself. In the frame of GIERE's perspectivism [108], the definition of model includes the purpose of use but also the agent who aims to use it. Therefore a minimal explication of model behavior and exploration of parameter roles is highly advised to decrease chances of misuses or misinterpretations of it. It includes simple runtime charts that are immediate on the NetLogo platform, but also indicators computations to evaluate outputs of the model. It can also be improved visualizations during runtime and model exploration, such as showed in Fig. 1.

2.1.2 *Sur le besoin d'exactitude dans l'implémentation du modèle*

Possible divergences between model description in a paper and the effectively implemented processes may have grave consequences on the reproducibility of science. The road network growth model given in [19] is one example that we are currently investigating. A strict implementation of model mechanisms provide slightly different results than the one presented in the paper, and as source code is not provided we need to test different hypotheses on possible mechanisms added by the programmer (that seems to be a connexion rule to intersections under a certain distance threshold). Lessons that could be possibly drawn from this examples are

- the necessity of providing source code
- the necessity of providing architecture description along with code (if model description is in a langage too far from architectural specifications) in order to identify possible implementation biaises
- the necessity of performing and detailing explicitly model explorations, that would in that case have helped to identify the implementation bias.

The last point, if first not provided, may ensure a limited risk of scientific falsification as it may be more complicated to fake false exploration results than to effectively explore the model. A joint project currently done is the writing of a false modeling paper in the spirit of [147], in which opposite results to the effective results of a given model are provided, without providing model implementation.

A first bunch of test is the acceptance of a clearly non-reproducible paper in diverse journals, possibly with a control on textual elements (using or not “buzz-words” associated to the journal, etc.). Depending on results, a second experiment may be tested with providing open source code for model implementation but still with false results, to verify if reviewers effectively try to reproduce results when they pretend to want the code (in reasonable computational power limits of course, HPC being not currently broadly available in Humanities).

2.1.3 *Perspectives*

Again, reproducibility and transparency is a non-negotiable feature of contemporaneous science, along with Open practices and Open Access. Too much examples (see a very recent one in experimental economics [53]) show in various disciplines the lack of reproducibility of experiments, that is a falsification of previous results or a result in itself. Falsification is a costly practice, and even if necessary [60], could be made more efficient through more transparency and direct reproducibility, increase therein the global workflow of science. We develop in parallel of this thesis various tools aimed to ease reproducibility, for which an overview is given in appendix 17.

2.2 UN CADRE UNIFIÉ POUR LES MODÈLES STOCHASTIQUES DE CROISSANCE URBAINE

Urban growth modeling fall in the case of tentatives to find self-consistent rules reproducing dynamics of an urban system, and thus in our logic of system morphogenesis. We examine here methodological issues linked to different frameworks of urban growth.

2.2.1 *Introduction*

Various stochastic models aiming to reproduce population patterns on large temporal and spatial scales (city systems) have been discussed across various fields of the literature, from economics to geography, including models proposed by physicists. We propose here a general framework that allows to include different famous models (in particular Gibrat, Simon and Preferential Attachment model) within an unified vision. It brings first an insight into epistemological debates on the relevance of models. Furthermore, bridges between models lead to the possible transfer of analytical results to some models that are not directly tractable.

Seminal models of urban growth are Simon [236] (later generalized as e.g. [127]) and Gibrat models. Many examples can be given across disciplines. [29] give an equation-based dynamical model, whereas [105] solves a stationary model. [106] reviews urban growth approaches in economics. A model adapted from evolutive urban theory is solved in [101] and improves Gibrat models. The question of empirical scales at which it is consistent to study urban growth was also tackled in the particular case of France [47]. We stay to a certain level of tractability to include models as essence of our approach is links between models but do not make ontologic assumptions.

2.2.2 *Cadre de Travail*

What we propose as a framework can be understood as a meta-model in the sense of [72], i.e. an modular general modeling process within each model can be understood as a limit case or as a specific case of another model. More simply it should be a diagram of formal relations between models. The ontological aspect is also tackled by embedding the diagram into an ontological state space (which discretization corresponds to the “bricks” of the incremental construction of [72]). It constructs a sort of model classification or modelography.

We are still at the stage of different derivations of links between models that are presented hereafter.

2.2.3 Dérivations

Généralisation de l'Attachement Préférentiel

[269] give a generalization of the classical Preferential Attachment Network Growth model, as a birth and death model with evolving entities. More precisely, network units gain and lose population (equivalent to links connexions) at fixed probabilities, and new unit can be created at a fixed rate.

Lien entre Gibrat et Attachement Préférentiel

Considérons un modèle de croissance strictement positive de Gibrat donnée par $P_i(t) = R_i(t) \cdot P_i(t-1)$ avec $R_i(t) > 1$, $\mu_i(t) = \mathbb{E}[R_i(t)]$ et $\sigma_i(t) = \mathbb{E}[R_i(t)^2]$. D'autre part, soit un modèle simple d'attachement préférentiel, avec une probabilité d'attachement $\lambda \in [0, 1]$ et un nombre de nouveau arrivants $m > 0$. Il est possible de dériver que le Gibrat est statistiquement équivalent à une limite de l'attachement préférentiel, sous l'hypothèse que toutes les fonctions génératrices des moments de $R_i(t)$ existent. Les distributions classiques qui peuvent être utilisées dans ce cas, e.g. une distribution normale ou log-normale, sont entièrement déterminées par leur deux premiers moments, ce qui rend cette hypothèse raisonnable.

Lemma 1 *The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.*

La preuve est donnée en Annexe 11.

Lien entre Simon et Attachement Préférentiel

A rewriting of Simon model yields a particular case of the generalized preferential attachment, in particular by vanishing death probability.

Lien entre Favaro-Pumain et Gibrat

[101] generalizes Gibrat models with innovation propagation dynamics, being therefore a generalization of that model. Theoretically, a process-based model equivalent to the Favaro-pumain should then fill the missing case in model classification at the corresponding discretization. Simpop models do not fill that case as they stay at the scale of city systems, as for Marius models [70]. These must also have their counterparts in discrete microscopic formulation.

Lien entre Bettencourt-West et Pumain

We are considering to study Bettencourt-West model for urban scaling laws [33] as entering the stochastic urban growth framework as stationary component of a random growth model, but investigation are still ongoing.

Autres modèles

[105] develops an economic model giving a Simon equivalent formulation. They in particular find out that in upper tail, proportional growth process occurs. We find the same result as a consequence of the derivation of the link between Gibrat and Preferential attachment models.

2.3 SENSIBILITÉ DES LOIS D'ECHELLE URBAINES À L'ETENDUE SPATIALE

Au centre de la théorie évolutive des villes se trouvent la hiérarchie et les lois d'échelle associées. Nous proposons ici un bref développement méthodologique sur la sensibilité des lois d'échelle à la définition de la ville.

Les lois d'échelle ont été montrées universelles des systèmes urbains à de nombreuses échelles et pour différents indicateurs. Des études récentes questionnent toutefois la cohérence de la détermination des exposants d'échelle, puisque leur valeur peut varier significativement selon les seuils utilisés pour définir les entités urbaines sur lesquelles les quantités urbaines sont intégrées, franchissant même dans certains cas la barrière qualitative de l'échelle linéaire, d'une loi infra-linéaire à une loi super-linéaire. Nous utilisons un modèle théorique simple de distribution spatiale des densités et des fonctions urbaines pour montrer analytiquement qu'un tel comportement peut être dérivé comme conséquence du type de distribution spatiale et de la méthode utilisée. Les simulations numériques confirment les résultats théoriques et révèle que les résultats sont raisonnablement indépendants du noyau spatial utilisé pour distribuer la densité.

Les lois d'échelle pour les systèmes urbains, en commençant par la bien connue loi rang-taille de Zipf pour la distribution des tailles des villes [105], ont été montrées être une caractéristique récurrente des systèmes urbains, à différentes échelles et pour différents types d'indicateurs. Elles reposent sur la constatation empirique que des indicateurs calculés sur des éléments du système urbain, qui peuvent être les villes dans le cas d'un système de villes, mais aussi des entités plus petites à une plus petite échelle, suivent relativement bien une distribution en loi de puissance en fonction de la taille de l'entité, i.e. pour l'entité i avec population P_i , on a pour une quantité intégrée A_i , la relation $A_i \simeq A_0 \cdot \left(\frac{P_i}{P_0}\right)^\alpha$. Les exposants d'échelle α peuvent être plus petits ou plus grands que 1, menant à des effets infra ou supra-linéaires. Diverses interprétations thématiques de ce phénomène ont été proposées, typiquement sous la forme d'analyse des processus. La littérature économique contient une production abondante sur le sujet (voir [106] pour une revue), mais est généralement faiblement spatiale, donc de faible intérêt pour notre approche qui s'intéresse particulièrement à l'organisation spatiale. Des règles économiques simples comme un équilibre énergétique peut conduire à de simples lois d'échelles [33] mais sont difficiles à ajuster empiriquement. Une proposition intéressante par PUMAIN est qu'elles sont intrinsèquement dues au caractère évolutionnaire des systèmes de villes, où l'émergence complexe par les interactions entre villes génère de telles distributions globales [209]. Même si un parallèle tenant peut être fait avec les systèmes biologiques auto-organisés, Pu-

MAIN insiste sur le fait que l'hypothèse d'ergodicité pour de tels systèmes n'est pas raisonnable dans le cas de système géographiques et que l'analogie est difficilement exploitable [207]. D'autres explications ont été proposées à d'autres échelles, comme le modèle de croissance urbaine à échelle mesoscopique (échelle de la ville) donné dans [163] qui montre que la congestion dans les réseaux de transport pourrait être une raison de la forme des villes et des lois d'échelle correspondantes. On peut noter que les modèles "classiques" de croissance urbaine comme le modèle de Gibrat [101] fournissent une approximation au premier ordre des systèmes exhibant des lois d'échelles, mais que les interactions entre agents doivent être incorporées dans le modèle pour obtenir un résultat plus fidèle aux données réelles, comme le modèle de Favaro-Pumain pour la propagation des cycles d'innovation proposé dans [101], qui généralise un modèle de Gibrat pour la croissance des villes françaises avec une ontologie similaire à celle des modèles Simpop.

The derivations in the simple case of exponential mixture density, are done in Appendix 11.

2.4 CONTRÔLE STATISTIQUE POUR LES CONDITIONS INITIALES PAR GÉNÉRATION DE DONNÉES SYNTHÉTIQUES

2.4.1 Contexte

When evaluating data-driven models, or even more simple partially data-driven models involving simplified parametrization, an unavoidable issue is the lack of control on “underlying system parameters” (what is a ill-defined notion but should be seen in our sense as parameters governing system dynamics). Indeed, a statistics extracted from running the model on enough different datasets can become strongly biased by the presence of confounding in the underlying real data, as it is impossible to know if result is due to processes the model tries to translate or to a hidden structure common to all data.

We formalize briefly a proposition of method that would allow to add controls on meta-parameters, in the sense of parameters driving the represented system at a higher temporal and spatial scale, for a model of simulation. We make the hypothesis that such method is valid under constraints of disjunction for scales and/or ontologies between the model of simulation and the domain of meta-parameters.

2.4.2 Description

An advanced knowledge of the behavior of computational models on their parameter space is a necessary condition for deductions of thematic conclusions or their practical application [12]. But the choice of varying parameters is always subjective, as some may be fixed by a real-world parametrization, or other may be interpreted as arbitrarily fixed initial conditions. It raises methodological and epistemological issues for the sensitivity analysis, as the scope of the model may become ill-defined.

Let consider the concrete example of the Schelling Segregation model [232]. One of its crucial features on which the literature has been rather controversial is the influence of the spatial structure of the container on which agents evolve. The thematic aim of the project developed in [74] is to clarify this point through a systematic model exploration. A methodological contribution is the construction of a framework allowing the analysis of the sensitivity of models to *meta-parameters*, i.e. to parameters considered as fixed initial conditions (e.g. the spatial structure for the Schelling model), or to parameters of another model generating an initial configuration yielding thus a *simple coupling* between models (serial coupling). The benefits of such an approach are various but include for example the knowledge of model behavior in an extended frame, the possibility of statistical control when regressing model outputs, a finer exploration of model

derivatives than with a naive approach. Some remarks can be made on the approach :

- What knowledge are brought by adding the upstream model, rather than for example in the Schelling case exploring a large set of initial geometries ?

→ *to obtain a sufficiently large set of initial configuration, one quickly needs a model to generate them; in that case a quasi-random generation followed by a filtering on morphological constraint will be a morphogenesis model, which parameters are the ones of the generation and the filtering methods. Furthermore, as detailed further, the determination of the derivative of the downstream model is made possible by the coupling and knowledge of the upstream model.*
- Statistical noise is added by coupling models

→ *Repetitions needed for convergence are indeed larger as the final expectance has to be determined by repeating on the first times the second model; but it is exactly the same as exploring directly many configuration, to obtain statistical robustness in that case one must repeat on similar configurations.*
- Complexity is added by coupling models

→ *In the sense of Varenne [250], coupling is simple and no complexity is thus added.*

2.4.3 Description Formelle

One has the composition of the derivative along the meta-parameter

$$\partial_\alpha [M_u \circ M_d] = (\partial_\alpha M_u \circ M_d) \cdot \partial_\alpha M_d$$

→ *the sensitivity of the downstream model (Schelling) can be determined by studying the serial coupling and the upstream model; thematic knowledge : sensitivity to an implicit meta-parameter; and computational gain : generation of controlled differentiates in the “initial space” is quasi impossible.*

The question of stochasticity in simply coupled models causes no additional issue as $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. It naturally multiplies the number of repetition needed for convergence what is the expected behavior.

2.5 LIEN ENTRE CORRELATION SPATIO-TEMPORELLES STATIQUES ET DYNAMIQUES SOUS HYPOTHÈSES SIMPLIFIÉES

Space and Time are both crucial for the study of geographical systems when aiming to understand *processes* (by definition dynamical [135]) evolving in a *spatial structure* in the sense of [92]. Space is more than coordinates for elements of the system, but a dimension in itself that drives interactions and thus system properties. Reading geographical systems from the point of view of *spatio-temporal processes* emphasizes the fact that *space actually matters*. Space and time are closely linked in such processes, and depending on the underlying mechanisms, one can expect to extract useful information from one on the other : in certain cases that we will investigate in this part, it is for example possible to learn about dynamics from static information. Spatio-temporal correlations approaches, linked to spatio-temporal dynamics, are present in very broad fields such as dynamical image processing (including video compression) [55, 126, 140], target tracking [28, 254], climate science [76], Earth sciences [170], city systems dynamics [128, 196], among others.

[77] : spatio-temporal chaos

The capture of neighborhood effects in statistical models is a wisely used practice in spatial statistics, as the technique of Geographically Weighted Regression illustrates [51]. A possible interpretation among many definitions of spatial autocorrelation [115] yields that by estimating a plausible characteristic distance for spatial correlations or auto-correlations, one can isolate independent effects between variables from effects due to neighborhood interactions³. The study of the spatial covariance structure is a cornerstone of advanced spatial statistics that was early formulated [114]. We propose now to study possible links between spatial and temporal correlations, using spatio-temporal covariance structure to infer information on dynamical processes.

2.5.1 Notations

We consider a multivariate spatio-temporal stochastic process denoted by $\vec{Y}[\vec{x}, t]$. At a given point \vec{x}_0 in space, we can define temporal covariance structure by

$$\mathbf{C}_t(\vec{x}_0) = \text{Var}[\vec{Y}[\vec{x}_0, \cdot]]$$

and spatial covariance structure at fixed time by

$$\mathbf{C}_x(t) = \text{Var}[\vec{Y}[\cdot, t]]$$

³ note that the formal link between models of spatial autocorrelation (see e.g. [116]) is not clear and should be further investigated

It is clear that these quantities will be in practice first ill-defined because of the difficulty in interpreting such a process by a spatio-temporal random variable, secondly highly non-stationary in space and time. We stay however at a theoretical level to gain structural knowledge, reviewing simple cases in which a formal link can be established.

2.5.2 *Equation des Ondes*

In the case of propagating waves, there is an immediate link. Let assume that a wave equation if verified by “deterministic” parts of components

$$c^2 \cdot \partial_t^2 \bar{Y}_i = \Delta \bar{Y}_i \quad (1)$$

with $Y_i = \bar{Y}_i + \varepsilon_i$. If errors are uncorrelated and processes are stationary, we have then directly

$$\mathbf{C}_t [\partial_t^2 Y_i, \partial_t^2 Y_j] = \frac{1}{c^2} \cdot \mathbf{C}_x [\Delta Y_i, \Delta Y_j] \quad (2)$$

This gives us however few insight on real systems as local diffusion, stationary assumptions and uncorrelated noises are far from being verified in empirical situations.

2.5.3 *Equation de Fokker-Planck*

An other interesting approach may when the process verifies a Fokker-Planck equation on probabilities of the state of the system when it is given by its position (diffusion of particles in that case)

$$\partial_t P(x_i, t) = -d \cdot \partial_x P(x_i, t) + \frac{\sigma^2}{2} \partial_x^2 P(x_i, t) \quad (3)$$

With no cross-correlation terms in the Fokker-Planck equation, covariance between processes vanish. We have finally in that case only a relation between averaged spatial and temporal variances that brings no information to our question.

2.5.4 *Equation Maitresse*

In the case of a master equation on probabilities of discrete states of the system

$$\partial_t \vec{P} = \mathbb{W} \vec{P} \quad (4)$$

we have then for state i , $\partial_t P_i = \sum_j W_{ij} P_j$. As this relation is at a fixed time we can average in time to obtain an equation on temporal covariance. It is not clear how to make the link with spatial covariance as these will depend on spatial specification of discrete states. This question is still under investigation.

2.5.5 Echantillonnage spatial cohérent

In a more empirical way, we propose to not assume any constraint of process dynamics but to however investigate how the computation of spatial correlations can inform on temporal correlations. We try to formulate easily verifiable assumptions under which this is possible.

We make the following assumptions on the spatio-temporal stochastic processes $Y_i[\vec{x}, t]$:

1. Local spatial autocorrelation is present and bounded by l_ρ (in other words the processes are continuous in space) : at any \vec{x} and t , $|\rho_{\|\Delta\vec{x}\| < l_\rho} [Y_i(\vec{x} + \Delta\vec{x}, t), Y_i(\vec{x}, t)]| > 0$.
2. Processes are locally parametrized : $Y_i = Y_i[\alpha_i]$, where $\alpha_i(\vec{x})$ varies with l_α , with $l_\alpha \gg l_\rho$.
3. Spatial correlations between processes have a sense at an intermediate scale l such that $l_\alpha \gg l \gg l_\rho$.
4. Processes covariance stationarity times scale as \sqrt{l} .
5. Local ergodicity is present at scale l and dynamics are locally chaotic.

Assumptions one to three can be tested empirically and allow to compare spatial correlation estimated on spatial samplings at scale l . Assumption four is more delicate as we are precisely constructing this methodology because we have no temporal information on processes. It is however typical of spatial diffusion processes, and population or innovation diffusion should verify this assumption. The last assumption can be tested if feasible space is known, by checking cribbing on image space on the spatial sample. Under these conditions, local spatial sampling is equivalent to temporal sampling and spatial correlation estimators provide estimator of temporal correlations.

2.6 GÉNÉRATION DE DONNÉES SYNTHÉTIQUES CORRÉLÉES

La génération de données synthétiques hybrides similaires à des données réelles présente des enjeux méthodologiques et thématiques pour la plupart des disciplines dont l'objet est l'étude de systèmes complexes. Comme l'interdépendance entre les éléments constitutifs d'un système, matérialisée par leur relations, conduit à l'émergence de ses propriétés macroscopiques, une possibilité de contrôle de l'intensité des dépendances dans un jeu de données synthétiques est un instrument de connaissance du comportement du système. Nous proposons une méthodologie de génération de données synthétiques hybrides sur lequel la structure de correlation est contrôlée. La méthode est illustrée sur des séries temporelles financières et permet l'étude de l'interférence entre composantes à différentes fréquences sur la performance d'un modèle prédictif, en fonction des correlations entre composantes à différentes échelles. On présente ensuite une application à un système géographique, dans laquelle le couplage faible d'un modèle de distribution de densité de population avec un modèle de génération de réseau permet la simulation de configurations territoriales, qui sont calibrées selon des objectifs morphologiques sur l'ensemble de l'Europe. L'exploration intensive du modèle permet l'obtention d'un large spectre de valeurs pour la matrice de correlation entre mesures morphologiques et mesures du réseau. On démontre ainsi les possibilités d'applications variées et les potentialités de la méthode.

2.6.1 Contexte

L'utilisation de données synthétiques, au sens de populations statistiques d'individus générées aléatoirement sous la contrainte de reproduire certaines caractéristiques du système étudié, est une pratique méthodologique largement répandue dans de nombreuses disciplines, et particulièrement pour des problématiques liées aux systèmes complexes, telles que par exemple l'évaluation thérapeutique [1], l'étude des systèmes territoriaux [181, 201], l'apprentissage statistique [36] ou la bio-informatique [52]. Il peut s'agir d'une désagrégation par création d'une population au niveau microscopique présentant des caractéristiques macroscopiques données, ou bien de la création de nouvelles populations au même niveau d'agrégation qu'un échantillon donné avec un critère de ressemblance aux données réelles. Le niveau de ce critère peut dépendre des applications attendues et peut par exemple aller de la fidélité des distributions statistiques pour un certain nombre d'indicateurs à des contraintes plus faibles de valeurs pour des indicateurs agrégés, c'est à dire l'existence de motifs macroscopiques similaires. Dans le cas de systèmes chaotiques ou présentant de fortes caractéristiques d'émergence, une contrainte microscopique

pique n'implique pas nécessairement le respect des motifs macroscopiques, et arriver à les reproduire est justement un des enjeux des pratiques de modélisation et simulation en sciences de la complexité. La donnée, qu'elle soit simulée, mesurée ou hybride est au cœur de l'étude des systèmes complexes de par la maturation de nouvelles approches computationnelles [9], il est donc essentiel d'étudier des procédures d'extraction d'information des données (fouille de données) et de simulation d'une information similaire (génération de données synthétiques).

Si le premier ordre est de manière générale bien maîtrisé, il n'est pas systématique ni aisément de contrôler le second ordre, c'est à dire les structures de covariance entre les variables générées, même si des exemples spécifiques existent, comme dans [271] où la sensibilité des sorties de modèles de choix discrets à la forme des distributions des variables aléatoires ainsi qu'à leur structures de dépendance. Il est également possible d'interpréter les modèles de génération de réseaux complexes [186] comme la création d'une structure d'interdépendance au sein d'un système, représentée par la topologie des liens. Nous proposons ici une méthode générique prenant en compte l'interdépendance lors de la génération de données synthétiques, sous la forme de correlations.

L'ensemble des méthodologies mentionnées en introduction sont trop variées pour être résumées par un même formalisme. Nous proposons ici une formulation générique ne dépendant pas du domaine d'application, ciblée sur le contrôle de la structure de correlation des données synthétiques.

2.6.2 Formalisation

Soit un processus stochastique multidimensionnel \vec{X}_I (l'ensemble d'indexation pouvant être par exemple le temps dans le cas de séries temporelles, l'espace, ou une indexation quelconque). On se propose, à partir d'un jeu de réalisations $X = (X_{i,j})$, de générer une population statistique $\tilde{X} = \tilde{X}_{i,j}$ telle que

1. d'une part un certain critère de proximité aux données est vérifié, i.e. étant donné une précision ϵ et un indicateur f , $\|f(X) - f(\tilde{X})\| < \epsilon$
2. d'autre part le niveau de correlation est contrôlé, i.e. étant donné une matrice fixant une structure de covariance R , $\text{Var}[(\tilde{X}_i)] = R$, où la matrice de variance/covariance est estimée sur la population synthétique.

La satisfaction du deuxième point sera généralement conditionnée par la valeur de paramètres, dont dépendra la procédure de génération, qu'il s'agisse de modèles simples ou complexes. Formellement, les processus synthétiques sont des familles paramétriques

$\tilde{X}_i[\vec{\alpha}]$. Nous proposons de décliner cette méthode sur deux exemples très différents mais tous deux typiques des systèmes complexes : des séries temporelles financières à haute fréquence, et les systèmes territoriaux. On illustre ainsi la flexibilité de la logique, ouvrant des portes interdisciplinaires par l'exportation de méthodes ou raisonnements par exemple. Dans le premier cas, la proximité aux données est l'égalité des signaux à une fréquence fondamentale, auxquels on superpose des composantes synthétiques dont il est facile de contrôler le niveau de correlation. On se place dans une logique de données hybrides, pour tester des hypothèses ou modèles dans un contexte plus proche de la réalité que sur des données purement synthétiques. Cet exemple, sans rapport thématique avec la thèse, est présenté en Appendice 12. Dans le deuxième cas, la calibration morphologique d'un modèle de distribution de densité de peuplement permet de respecter le critère de proximité aux données. Les correlations de la forme urbaine avec celle d'un réseau de transport sont ensuite obtenues empiriquement par exploration du couplage avec un modèle de génération de réseau. Leur contrôle est dans ce cas indirect puisque constaté empiriquement.

2.7 POUR UN USAGE RAISONNÉ DES DONNÉES MASSIVES ET DE LA COMPUTATION

La soi-disante *révolution des données massives* réside autant dans la disponibilité de grands jeux de données de nouveaux types variés, que dans la puissance de calcul potentielle toujours en augmentation. Même si le *tournant computationnel* ([9]) est central pour une science consciente de la complexité et est sans douter la base des pratiques de modélisation futures en géographie comme [12] souligne, nous soutenons que à la fois le *déluge de données* et les *capacités de calcul* sont dangereuses si non cadrées dans un cadre théorique et formel propre. Le premier peut biaiser les directions de recherche vers les jeux de données disponibles (comme par exemple les nombreuses étude de mobilité se basant sur twitter) avec le risque de se déconnecter d'un fond théorique, tandis que le second peut occulter des résolutions analytiques préliminaires essentielles pour un usage cohérent des simulations. Nous avançons que les conditions pour la majorité des résultats dans cette thèse sont en effet ceux mis en danger par un enthousiasme inconsidéré pour les données massives, tirant la conclusion qu'un challenge majeur pour la géocomputation future est une intégration sage des nouvelles pratiques au sein du corpus existant de connaissances.

La puissance de calcul disponible semble suivre un tendance exponentielle, comme une sorte de loi de Moore. Grace à d'une part la loi de Moore effective pour le matériel, d'autre part l'amélioration des logiciels et algorithmes, conjointement avec une démocratisation de l'accès au infrastructures de simulation à grande échelle, permet à toujours plus de temps processeur d'être disponible pour le chercheur en sciences sociales (et pour le scientifique en général, mais cette mutation a déjà été opérée depuis plus longtemps dans d'autres domaines, puisque par exemple le CERN est à la pointe en terme de calcul distant et sur grille). Il y a environ une dizaine d'année, [111] était forcé de conclure que les analyses de réseau, pour les transports publics parisiens, étaient "limitées par le calcul". Aujourd'hui la plupart des mêmes analyses seraient rapidement réglée sur un ordinateur personnel avec les logiciels et programmes appropriés : [150] est un témoin d'un tel progrès, introduisant des nouveaux indicateurs avec une plus grande complexité de calcul, qui sont calculés sur des réseaux à grande échelle. Le même parallèle peut être fait pour les modèles Simpop : les premiers modèles Simpop au début du millénaire [230] étaient "calibrés" à la main, tandis que [73] calibre le modèle Marius en multi-modélisation et [234] calibre très précisément le modèle SimpopLocal, chacun sur la grille avec des milliards de simulations. Un dernier exemple, le champ de la *Space Syntax*, a témoigné d'une longue route et de progrès considérables depuis ses

origines théoriques [131] jusqu'à ses récentes applications à grande échelle [130].

Concernant les nouvelles données "massives" qui sont disponibles, il est clair que des quantités toujours plus grandes et des types toujours nouveaux sont disponibles. De nombreux exemples de champs d'application peuvent être donnés. La mobilité en est typique, puisque étudiée selon divers points de vue, comme les nouvelles données issues des systèmes de transport intelligents [189], des réseaux sociaux [103], ou des données plus exotiques comme des données de téléphonie mobile [82]. Dans un autre esprit, l'ouverture de jeux de données "classiques" (comme les applications synthétiques urbaines, les initiatives gouvernementales pour les données ouvertes) devrait pouvoir toujours plus de méta-analyses. De nouvelles façons de pratiquer la recherche et produire des données sont également en train d'émerger, vers des initiatives plus interactives et venant de l'utilisateur. Ainsi, [69] décrit une application web ayant pour but de présenter une méta-analyse de la loi de Zipf sur de nombreux jeux de données, mais en particulier inclut une option de dépôt, à travers laquelle l'utilisateur peut télécharger son propre jeu de données et l'inclure dans la méta-analyse. D'autres applications permettent l'exploration interactive de la littérature scientifique pour une meilleure connaissance d'un horizon scientifique complexe, comme [58] fait.

Comme toujours la situation n'est naturellement pas aussi idyllique qu'elle semble être au premier abord, et l'herbe verte du pré du voisin que nous pouvons être tentés d'aller brouter se transforme rapidement en un triste fumier. En effet, les objectifs et motivations sont flous et on peut facilement s'y perdre. Des illustrations parleront d'elles-même. [21] introduit un nouveau jeu de données et des méthodes relativement nouvelles pour quantifier l'évolution du réseau de rues, mais les résultats, sur lesquels les auteurs semblent s'étonner, sont qu'une transition a eu lieu à Paris à l'époque d'Haussmann. Tout historien de l'urbanisme s'interrogerait sur le but exact de l'étude, puisque à la fin un sentiment étrange de réinvention de la roue flotte dans l'air. L'utilisation des ressources de calcul peut également être exagéré, et dans le cas de la modélisation multi-agent, on peut citer [10], pour lequel l'objectif de simuler le système à l'échelle 1 : 1 semble être loin des motivations et justifications originelles de la modélisation agent, et pourrait même donner des arguments aux économistes *mainstream* qui dénigrent facilement les ABMS. D'autres anecdotes peuvent inquiéter : [79] est une application web qui gâche des ressources de calcul pour simuler des distributions Gaussiennes afin de calculer pour un modèle de Gibrat, afin de calculer leur moyenne et variance, qui sont des paramètres d'entrée du modèle. En résumé, cela revient à vérifier le Théorème de la Limite Centrale, qui est a priori assez accepté par la plupart des scientifiques. D'autre part, la distribution complète donnée par un modèle de Gibrat est entière-

rement connue théoriquement comme résolu e.g. par [105]. Récemment, sur la liste de diffusion de géographie francophone *Geotamtam*, un soudain engouement autour des données issues de *Pokemon Go* a semblé répondre plus à un besoin urgent et inexplicable d'exploiter cette source de données avant tous les autres, plutôt qu'à des considérations théoriques élaborées. Des jeux de données existant et précis, comme la population historique des villes (pour la France la base Pumain-INED par exemple), sont loin d'être entièrement exploités et il pourrait être plus pertinent de se concentrer sur ces jeux de données classiques qui existent déjà. De même, il faut être conscient des possibles applications de résultats basée sur des malentendus : [162] fait une très bonne analyse de la redistribution potentielle des transactions de carte bancaire au sein d'une ville, mais présente les résultats comme la base possible de recommandations de politiques pour une équité sociale en agissant sur la mobilité, oubliant que la forme et les fonctions urbaines sont couplés de manière complexe et que déplacer des transactions d'un endroit à un autre implique des processus bien plus complexes que des régulations directes (même dans un pays comme la Chine où les régulations sont effectivement mise en place et imposées d'une main de fer).

Notre principal argument est que le tournant computationnel et les pratiques de simulation seront centrales en géographie, mais peuvent également être dangereux, pour les raisons illustrées ci-dessus, i.e. que le déluge de données peut imposer les sujets de recherche et occulter la théorie, et que la computation peut éluder la construction et la résolution de modèles. Un lien plus fort est nécessaire entre les pratiques de calcul, l'informatique, les mathématiques, les statistiques et la géographie théorique. La Géographie Théorique et Quantitative est au centre de cette dynamique, puisqu'il s'agit

2.8 UN CADRE BASÉ SUR LA DISCRÉPANCE POUR COMPARER LA ROBUSTESSE DES EVALUATIONS MULTI-ATTRIBUTS

Les évaluations multi-objectifs sont un aspect essentiel de la gestion de systèmes complexes, puisque la complexité intrinsèque d'un système est généralement étroitement liée au nombre d'objectifs d'optimisation potentiels. Cependant, une évaluation ne fait pas sens si sa robustesse, au sens de sa fiabilité, n'est pas donnée. Les méthodes statistiques usuelles fournissant une mesure de robustesse sont très dépendantes des modèles sous-jacents. Nous proposons une formulation d'un cadre indépendant du modèle, dans le cas d'indicateurs intégrés et agrégés (évaluation multi-attributs), qui permet de définir une mesure de robustesse relative prenant en compte la structure des données et les valeurs des indicateurs. La méthode est testée sur données urbaines synthétiques associées aux arrondissements de Paris, et à des données réelles de revenus pour l'évaluation de la ségrégation urbaine dans la région métropolitaine du Grand Paris. Les premiers résultats numériques montrent les potentialités de cette nouvelle méthode. De plus, sa relative indépendance au type de système et au modèle pourrait la positionner comme une alternative aux méthodes statistiques classiques d'évaluation de la robustesse.

2.8.1 *Introduction*

Contexte Général

Les problèmes multi-objectifs sont organiquement liés à la complexité des systèmes sous-jacents. En effet, que ce soit dans le champ des *Systèmes Complexes Industriels*, dans le sens de systèmes conçus par ingénierie, où la construction de Systèmes de Systèmes (SoS) par couplage et intégration induit souvent des objectifs contradictoires [179], ou dans le champ des *Systèmes Complexes Naturels*, au sens de systèmes non désignés, physiques, biologiques ou sociaux, qui présentent des propriétés d'émergence et d'auto-organisation, pour lesquels les objectifs peuvent e.g. être le résultat de l'interaction d'agents hétérogènes (voir [185] pour une revue étendue des types de systèmes concernés par cette approche), l'optimisation multi-objectifs peut être explicitement introduite pour étudier ou désigner le système, mais régit généralement déjà implicitement les mécanismes internes du système. Le cas des Systèmes Complexes Sociaux-techniques est particulièrement intéressant puisque selon Haken [122], ils peuvent être vus comme des systèmes hybrides embarquant des agents sociaux dans des "artefacts techniques" (parfois jusqu'à un niveau inattendu, créant ce que PICON décrit comme *cyborgs* [195]), et cumulent ainsi la potentialité d'être à l'origine de problèmes multi-objectifs⁴. La no-

⁴ Nous désignons ici par *Evaluation Multi-objectifs* toutes les pratiques incluant le calcul de multiples indicateurs d'un système (il peut s'agir d'optimisation multi-objectif

tion récente d'*éco-quartier* [237] est un exemple typique pour lequel la durabilité implique des objectifs contradictoires. L'exemple des systèmes de transport, dont la conception a glissé durant la seconde moitié du 20ème siècle d'analyses coût-bénéfices à la price de décision multi-critères, est également typique de tels systèmes [25]. Les systèmes géographiques sont à présent bien étudiés d'un tel point de vue, en particulier grâce à l'intégration des cadres multi-objectifs au sein des Systèmes d'Information Géographiques [54]. Comme dans le cas microscopique des éco-quartiers, la planification et le design urbains mésoscopiques et macroscopiques peuvent être rendus durables grâce aux évaluations par indicateurs [138].

Un aspect crucial de l'évaluation est une certaine notion de sa fiabilité, que nous nommerons ici *robustesse*. Les méthodes statistiques incluent naturellement cette notion puisque la construction et l'estimation de modèles statistiques donne divers indicateurs de la consistance des résultats [152]. Le premier exemple venant à l'esprit est l'application de la loi des grands nombres pour obtenir la *p-valeur* d'une estimation de modèle, qui peut être interprété comme une mesure de confiance en les valeurs estimées. D'autre part, les intervalles de confiance et le *beta-power* sont d'autres indicateurs importants de robustesse statistique. L'inférence bayésienne fournit également des mesures de robustesse quand la distribution des paramètres est estimée de manière séquentielle. Concernant les optimisations multi-objectifs, en particulier par des algorithmes heuristiques (comme par exemple les algorithmes génétiques, ou les solveurs de recherche opérationnelle), la notion de robustesse d'une solution consiste plus en la stabilité de la solution dans l'espace des phases du système dynamique correspondant. Des progrès récents ont été faits vers une formulation unifiée de la robustesse pour les problèmes d'optimisation multi-objectifs, comme dans [83] où les fronts de Pareto robustes sont définis comme des solutions insensibles aux petites perturbations. Dans [18], la notion de degré de robustesse est introduite, formalisée comme une sorte de continuité des autres solutions dans des voisinages successifs d'une solution.

Cependant, il n'existe pas de méthode générique qui permettrait une évaluation de la robustesse de façon indépendante au modèle, i.e. qui serait extraite de la structure des données et des indicateurs mais ne dépendrait pas de la méthode utilisée. Un avantage serait par exemple une estimation *a priori* de la robustesse potentielle d'une évaluation et de décider ainsi si elle vaut la peine d'être faite. Nous proposons un cadre répondant à cette contrainte dans le cas particulier des évaluations multi-attributs, i.e. quand le problème est rendu unidimensionnel par agrégation des objectifs. Il est basé sur les données et non sur les modèles, au sens où l'estimation de la robustesse

pour un design de système, une évaluation multi-objectif d'un système existant, une évaluation multi-attributs ; notre cadre particulier correspondra au dernier cas).

ne dépendra pas de la manière dont les indicateurs sont calculés, tant qu'ils respectent certaines hypothèses détaillées par la suite.

Approche Proposée

OBJECTIFS COMME INTÉGRALES SPATIALES Nous supposons que les objectifs peuvent être exprimés comme intégrales spatiales, ce qui devrait s'appliquer à tout système territorial, et nos cas d'application sont des systèmes urbains. Ce n'est pas si restrictif en terme d'indicateurs possibles si l'on utilise les bonnes variables et noyaux intégrés : de façon analogue à la méthode de Regression Géographique Pondérée [51], toute variable spatiale peut être intégrée contre des noyaux réguliers de taille variable et le résultats sera une agrégation spatiale dont la signification dépendra de l'étendue du noyau. Les exemples utilisés par la suite comme des moyennes conditionnelles ou des sommes vérifient parfaitement cette hypothèse. Même un indicateur déjà agrégé dans l'espace peut être interprété comme une intégrale spatiale en utilisant une distribution de Dirac au centroïde de la zone correspondante.

OBJECTIFS AGRÉGÉS LINÉAIREMENT Une seconde hypothèse que nous faisons est que l'évaluation multi-objectifs est effectuée par agrégation linéaire des objectifs, c'est à dire qu'on se place dans le cadre d'un problème d'optimisation multi-attributs. Si $(q_i(\vec{x}))_i$ sont les valeurs des fonctions objectifs, on définit alors des poids $(w_i)_i$ afin de construire la fonction de prise de décision $q(\vec{x}) = \sum_i w_i q_i(\vec{x})$, dont la valeur détermine ensuite la performance d'une solution. Cette approche est analogue aux utilités agrégées en économie et est utilisée dans de nombreux domaines. La subtilité réside dans le choix des poids, i.e. de la forme de la fonction de projection, et différentes solutions ont été développées pour obtenir des poids selon la nature du problème. Récemment, [91] a proposé de comparer la robustesse des différentes techniques d'agrégation par une analyse de sensibilité, effectuée par simulations de Monte-Carlo pour produire des données synthétiques, ce qui permet d'obtenir la distribution des biais pour les différentes techniques, certaines étant significativement plus performantes que d'autres. Toutefois, la quantification de la robustesse dépend toujours des modèles utilisés dans ce travail.

Le reste de cette monographie est organisé de la façon suivante : la section 2 décrit intuitivement puis mathématiquement le cadre proposé ; la section 3 détaille ensuite l'implémentation, la collecte des données pour les cas d'étude et les résultats numériques pour une évaluation intra-urbaine synthétique et un cas réel métropolitain ; la section 4 discute finalement les limitations et les potentialités de la méthode.

2.8.2 Description du Cadre

Description Intuitive

Nous décrivons à présent le cadre proposé pour permettre théoriquement de comparer la robustesse d'évaluation de deux systèmes urbains différents. Ce cadre est une généralisation d'une méthode empirique proposée dans [5] pour accompagner une étude dans un autre contexte effectuant une comparaison du sens et de la pertinence des indicateurs dans un contexte de durabilité. Intuitivement, la base empirique se base sur les principes suivants :

- Les systèmes urbains peuvent être vus selon l'information disponible, i.e. les données brutes décrivant le système. Dans une approche basée sur les données, celles-ci sont la base de notre cadre et la robustesse sera déterminée par leur structure.
- A partir des données sont capturés des indicateurs (fonctions objectifs). Nous supposons qu'un choix d'indicateurs est une intention particulière de traduire des aspects particuliers du système, i.e. de capturer une réalisation d'un "fait urbain" au sens de MANGIN [174] - une sorte de fait stylisé en terme de processus et de mécanismes, ayant différentes réalisations sur des systèmes distincts dans l'espace, dépendant de chaque contexte géographique précis.
- Etant donné plusieurs systèmes et indicateurs associés, un espace commun peut être construit pour les comparer. Dans cet espace, les données représentent plus ou moins bien le système réel, c'est à dire qu'elles sont imprécises en fonction de l'échelle initiale, de la précision effective des données. Nous proposons de capturer exactement ces différents aspects au travers de la notion de discrépance d'un nuage de points, qui est un outil mathématique provenant des théories d'échantillonnage, permettant d'exprimer la façon dont un jeu de données rempli l'espace dans lequel il s'insère [88].

Synthétisant ces contraintes, nous proposons une notion de *Robustesse* d'une évaluation qui capture à la fois, en combinant la fiabilité des données à l'importance relative des indicateurs,

1. *Données manquantes* : une évaluation se basant sur des jeux de données plus raffinés sera naturellement plus robuste.
2. *Importance des indicateurs* : les indicateurs avec plus d'importance relative pèsent plus dans la robustesse totale.

Description Formelle

INDICATEURS Soit $(S_i)_{1 \leq i \leq N}$ un nombre fini de systèmes territoriaux géographiquement disjoints, que nous supposons décrits par les données brutes et des indicateurs intermédiaires, donnés par $S_i = (X_i, Y_i) \in \mathcal{X}_i \times \mathcal{Y}_i$ avec $\mathcal{X}_i = \prod_k \mathcal{X}_{i,k}$ tel que chaque sous-espace contient des matrices réelles : $\mathcal{X}_{i,k} = \mathbb{R}^{n_{i,k}^X p_{i,k}^X}$ (de la même façon pour \mathcal{Y}_i). Nous définissons également une fonction d'indice ontologique $I_X(i, k)$ (resp. $I_Y(i, k)$) prenant des valeurs entières qui coincident si et seulement si les deux variables ont même ontologie au sens de [161], c'est à dire qu'elles sont supposées représenter le même objet réel. On distingue les "données brutes" X_i à partir desquelles les indicateurs sont calculés généralement par des fonctions déterministes explicites, des "indicateurs intermédiaires" Y_i qui sont déjà intégrés et peuvent être par exemple les sorties de modèles élaborés simulant certains aspects du système urbain. Nous définissons l'espace caractéristique du "fait urbain" par

$$(\mathcal{X}, \mathcal{Y}) \underset{\text{def}}{=} \left(\prod \tilde{\mathcal{X}}_c \right) \times \left(\prod \tilde{\mathcal{Y}}_c \right) = \left(\prod_{\mathcal{X}_{i,k} \in \mathcal{D}_X} \mathbb{R}^{p_{i,k}^X} \right) \times \left(\prod_{\mathcal{Y}_{i,k} \in \mathcal{D}_Y} \mathbb{R}^{p_{i,k}^Y} \right) \quad (5)$$

avec $\mathcal{D}_X = \{\mathcal{X}_{i,k} | I(i, k) \text{ distincts, } n_{i,k}^X \text{ maximal}\}$ (de même pour \mathcal{Y}_i). Il s'agit en fait de l'espace abstrait sur lequel les indicateurs sont intégrés. Les indices c introduit par définition correspondent aux différents indicateurs au sein des différents systèmes. Cette espace est l'espace minimal commun à tous les systèmes permettant une définition commune des indicateurs pour tous.

Soit $X_{i,c}$ les données projetées canoniquement sur le sous-espace correspondant, bien définies pour tout i et tout c . Nous faisons donc l'hypothèse clé que tous les indicateurs sont calculés par intégration contre un noyau donné, i.e. pour tout c il existe H_c espace de fonctions à valeurs réelles sur $(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)$, tel que pour tout $h \in H_c$:

1. h est "suffisamment" régulière (distribution tempérée par exemple)
2. $q_c = \int_{(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)} h$ est une fonction décrivant le "fait urbain" (l'indicateur en lui-même)

Des exemples typiques de noyaux peuvent être :

- Une moyenne des lignes de $X_{i,c}$ est calculée par $h(x) = x \cdot f_{i,c}(x)$ où $f_{i,c}$ est la densité de la distribution de la variable sous-jacente.
- Un taux d'éléments du jeu de données respectant une condition donnée C , $h(x) = f_{i,c}(x) \chi_{C(x)}$.

- Pour des variables déjà agrégées \mathbf{Y} , une distribution de Dirac permet de les exprimer également comme des intégrales de noyaux.

AGRÉGATION La détermination des poids est en fait le point crucial des processus de prise de décision multi-attributs, et de nombreuses méthodes sont disponibles (voir [255] pour une revue dans le cas particulier de la gestion de l'énergie durable). Définissons les poids pour l'agrégation linéaire. Nous supposons les indicateurs normalisés, i.e. $q_c \in [0, 1]$, pour une construction plus simple des poids relatifs. Pour i, c et $h_c \in H_c$ donnés, le poids $w_{i,c}$ est simplement constitué par l'importance relative de l'indicateur $w_{i,c}^L = \frac{\hat{q}_{i,c}}{\sum_c \hat{q}_{i,c}}$ où $\hat{q}_{i,c}$ est un estimateur de q_c pour les données $X_{i,c}$ (i.e. la valeur calculée effectivement). On peut noter que cette étape n'est pas contrainte et que cela peut être étendu à tout ensemble d'attribution de poids, en prenant par exemple $\tilde{w}_{i,c} = w_{i,c} \cdot w'_{i,c}$ si w' sont les poids fixés par le preneur de décisions. Nous nous concentrerons sur l'influence relative des attributs et pour cela choisissons cette forme simple pour les poids.

ESTIMATION DE LA ROBUSTESSE La scène est à présent prête pour permettre d'estimer la robustesse d'une évaluation faite par la fonction d'agrégation. Pour cela, nous appliquons un théorème d'approximation d'intégrale similaire au méthodes introduites dans [253], puisque la forme intégrée des indicateurs permet justement de bénéficier de tels résultats théoriquement puissant. Soit $\mathbf{X}_{i,c} = (\vec{X}_{i,c,l})_{1 \leq l \leq n_{i,c}}$ et $D_{i,c} = \text{Disc}_{\vec{X}_c, L^2}(\mathbf{X}_{i,c})$ le discrépance du jeu de données⁵ [187]. Avec $h \in H_c$, on a la borne supérieure sur l'erreur d'approximation de l'intégrale

$$\left\| \int h_c - \frac{1}{n_{i,c}} \sum_l h_c(\vec{X}_{i,c,l}) \right\| \leq K \cdot \|h_c\| \cdot D_{i,c}$$

où K est une constante indépendante des points de données et des fonctions objectifs. Cela donne directement

$$\left\| \int \sum w_{i,c} h_c - \frac{1}{n_{i,c}} \sum_l w_{i,c} h_c(\vec{X}_{i,c,l}) \right\| \leq K \sum_c |w_{i,c}| \|h_c\| \cdot D_{i,c}$$

En supposant l'erreur réalisée de manière raisonnable (scénario du "pire de cas" pour la connaissance de la valeur théorique de la fonc-

⁵ La discrépance est définie comme la norme-L2 de la discrépance locale qui est pour des points de données normalisés $\mathbf{X} = (x_{ij}) \in [0, 1]^d$, une fonction de $t \in [0, 1]^d$ comparant le nombre de points compris dans le volume de l'hypercube correspondant, donné par $\text{disc}(t) = \frac{1}{n} \sum_i \mathbb{1}_{\prod_j x_{ij} < t_j} - \prod_j t_j$. C'est une mesure de la manière dont le nuage de points couvre l'espace.

tion agrégée), nous prenons cette borne supérieure comme une approximation de sa magnitude. De plus, la normalisation des indicateurs implique que $\|\mathbf{h}_c\| = 1$. Nous proposons alors de comparer les bornes d'erreurs entre deux évaluations. Elle dépendent seulement de la distribution des données (équivalence à la *robustesse statistique*) et des indicateurs choisis (sorte de *robustesse ontologique*, i.e. est-ce que les indicateurs ont un sens réel dans le contexte choisi et est-ce que leur valeur fait sens), et sont un moyen de combiner ces deux types de robustesse dans une seule valeur.

Nous définissons ainsi un *ratio de robustesse* pour comparer la robustesse de deux évaluations par

$$R_{i,i'} = \frac{\sum_c w_{i,c} \cdot D_{i,c}}{\sum_c w_{i',c} \cdot D_{i',c}} \quad (6)$$

L'interprétation intuitive de cette définition est que l'on compare la robustesse des évaluations en comparant la plus grande erreur faite dans chaque cas selon la structure des données et l'importance relative.

En construisant une relation d'ordre sur les évaluations en comparant la position du ratio par rapport à un, il est clair qu'on obtient un ordre complet sur l'ensemble des évaluations possibles. Ce ratio devrait en théorie permettre de comparer n'importe quelle évaluation d'un système urbain. Afin de garder un sens ontologique à cela, il devrait être utilisé pour comparer des sous-systèmes disjoints avec une proportion raisonnable d'indicateurs en commun, ou le même sous-système avec des indicateurs différents. On peut noter que cela fournit un moyen de tester l'influence des indicateurs sur une évaluation, en analysant la sensibilité du ratio à leur suppression. Au contraire, la détermination d'un nombre "minimal" d'indicateurs faisant chacun varier le ratio fortement pourrait être un moyen d'isoler des paramètres essentiels régissant le sous-système.

2.8.3 Résultats

IMPLÉMENTATION Le pré-traitement des données géographiques est fait via QGIS [212] pour des raisons de performances. L'implémentation du coeur est faite en R [242] pour la flexibilité de la gestion des données et du traitement statistique. De plus, le package DiceDesign [102] conçu pour les expériences numériques et l'échantillonnage, permet un calcul efficient et direct des discordances. Enfin, tout aussi important, l'ensemble du code source est disponible de manière ouverte sur le dépôt git du projet⁶ pour permettre la reproductibilité et la réutilisation [220].

⁶ à <https://github.com/JusteRaimbault/RobustnessDiscrepancy>

Implémentation sur Données Synthétiques

Nous proposons dans un premier temps d'illustrer l'implémentation par une application à des données et indicateurs synthétiques, pour des indicateurs de qualité de vie intra-urbaine pour la ville de Paris.

COLLECTE DES DONNÉES Le cas virtuel se base sur des données géographiques réelle, en particulier pour les arrondissements parisiens. Nous utilisons les données disponibles par le projet OpenStreetMap [30] qui fournit déjà des données précises en haute définition pour de nombreux aspects urbains. Nous utilisons le réseau de rues et la position des bâtiments dans la ville de Paris. Les limites des arrondissements, utilisées pour agréger et extraire les features lorsqu'on travaille sur un seul district, sont aussi pris de la même source. Nous utilisons les centroïdes des polygones des bâtiments et les segments du réseau de rues. Le jeu de données brutes consiste d'environ 200k bâtiments et 100k segments de rues.

CAS VIRTUEL Nous travaillons sur chaque arrondissement de Paris (du 1er au 20ème) comme un système urbain évalué. Des données synthétiques aléatoires sont associées aux features spatiales, chaque arrondissement pouvant alors être évalué de manière stochastique, et des répétitions permettent d'obtenir le comportement statistique moyen des indicateurs jouets et des ratios de robustesse. Les indicateurs choisis doivent être calculés comme des indicateurs résidentiels et du réseau de rues. Pour montrer différents exemples, nous implementons deux kernels moyens et une moyenne conditionnelle, tous liés à la durabilité environnementale et la qualité de vie, chacun devant être maximisés. On peut noter que ces indicateurs ont un sens réel mais pas de raison particulière d'être agrégés, ils sont ici choisis pour l'aspect pratique du modèle jouet et de la génération de données synthétiques. Avec $a \in \{1 \dots 20\}$ le nombre d'arrondissements, $A(a)$ l'aire spatiale correspondante à chacun, $b \in B$ les coordonnées des bâtiments et $s \in S$ les segments de rues, nous prenons

- Le complémentaire de la distance journalière moyenne au travail en voiture par individu, approché par, avec $n_{cars}(b)$ nombre de voiture dans le bâtiment (généré aléatoirement en associant des voitures à bâtiments proportionnel au taux de motorisation attendu α_m 0.4 à Paris), d_w distance des individus à leur travail (généré à partir du bâtiment vers un point aléatoire distribué uniformément dans l'étendue spatiale du jeu de données), et d_{max} le diamètre de l'aire de Paris, $\bar{d}_w = 1 - \frac{1}{|b \in A(a)|} \cdot \sum_{b \in A(a)} n_{cars}(b) \cdot \frac{d_w}{d_{max}}$
- Le complémentaire des flots moyens de voitures des rues dans la zone, approché par, avec $\varphi(s)$ flot relatif dans le segment de rue s , généré par le minimum entre 1 et une distribution

log-normale ajustée pour avoir 95% de masse plus petite que 1, ce qui mimique la distribution hiérarchique de l'utilisation des rues (qui correspond à la centralité de chemin), et $l(s)$ longueur du segment, $\bar{\varphi} = 1 - \frac{1}{|s \in A(a)|} \cdot \sum_{s \in A(a)} \varphi(s) \cdot \frac{l(s)}{\max(l(s))}$

- Longueur relative de rues piétonnes \bar{p} , calculé via une dummy variable aléatoire uniforme ajustée pour obtenir une proportion fixée de segments pédestre.

Comme les données synthétiques sont stochastiques, les simulations sont lancées pour chaque quartier $N = 50$ fois, ce qui était un compromis raisonnable entre convergence statistique et temps nécessaire au calcul. La table 1 montre les résultats (moyennes et déviations standard) des valeurs des indicateurs et le calcul du ratio de robustesse. Les déviations standard obtenues confirment que ce nombre de simulations donnent des résultats consistants. Les indicateurs obtenus en fixant un ratio fixe montrent peu de variabilité, ce qui peut être une limite de cette approche jouet. On obtient toutefois le résultat intéressant que la majorité des arrondissements donne des évaluations plus robustes que le 1er arrondissement, ce qui était attendu par la taille et la fonction de ce quartier : il s'agit en effet d'un petit quartier avec de grand bâtiment administratifs, ce qui implique moins d'éléments spatiaux et pour cela une évaluation moins robuste selon la définition qu'on en a donnée.

Application à un cas réel : ségrégation métropolitaine

Le premier exemple avait pour but de montrer les potentialités de la méthode mais était purement synthétique, ne pouvant pour cela fournir pas de conclusion concrète ni d'implications pour la gouvernance. Nous proposons maintenant de l'appliquer à des données réelles dans le cas de la ségrégation métropolitaine.

DONNÉES Nous travaillons sur les données de revenus, disponible pour la France à un niveau intra-urbain (unités statistiques élémentaires IRIS) pour l'année 2011 sous la forme de résumé statistiques (déciles uniquement si la zone est peuplée suffisamment pour assurer l'anonymat), fournies par l'INSEE⁷. Les données sont associées à l'étendue géographique des unités statistiques, permettant le calcul d'indicateurs d'analyse spatiale.

INDICATEURS Nous utilisons ici trois indicateurs de ségrégation intégrés sur une zone géographique. Supposons la zone divisée en unités couvrantes S_i pour $1 \leq i \leq N$ avec pour centroïdes (x_i, y_i) . Chaque unité a des caractéristiques de population P_i et de revenu

⁷ <http://www.insee.fr>

Arrdt	$\langle \bar{d}_w \rangle \pm \sigma(\bar{d}_w)$	$\langle \bar{\varphi} \rangle \pm \sigma(\bar{\varphi})$	$\langle \bar{p} \rangle \pm \sigma(\bar{p})$	$R_{i,1}$
1 th	0.731655 ± 0.041099	0.917462 ± 0.026637	0.191615 ± 0.052142	1.000000 ± 0.000000
2 th	0.723225 ± 0.032539	0.844350 ± 0.036085	0.209467 ± 0.058675	1.002098 ± 0.039972
3 th	0.713716 ± 0.044789	0.797313 ± 0.057480	0.185541 ± 0.065089	0.999341 ± 0.048825
4 th	0.712394 ± 0.042897	0.861635 ± 0.030859	0.201236 ± 0.044395	0.973045 ± 0.036993
5 th	0.715557 ± 0.026328	0.894675 ± 0.020730	0.209965 ± 0.050093	0.963466 ± 0.040722
6 th	0.733249 ± 0.026890	0.875613 ± 0.029169	0.206690 ± 0.054850	0.990676 ± 0.031666
7 th	0.719775 ± 0.029072	0.891861 ± 0.026695	0.209265 ± 0.041337	0.966103 ± 0.037132
8 th	0.713602 ± 0.034423	0.931776 ± 0.015356	0.208923 ± 0.036814	0.973975 ± 0.033809
9 th	0.712441 ± 0.027587	0.910817 ± 0.015915	0.202283 ± 0.049044	0.971889 ± 0.035381
10 th	0.713072 ± 0.028918	0.881710 ± 0.021668	0.210118 ± 0.040435	0.991036 ± 0.038942
11 th	0.682905 ± 0.034225	0.875217 ± 0.019678	0.203195 ± 0.047049	0.949828 ± 0.035122
12 th	0.646328 ± 0.039668	0.920086 ± 0.019238	0.198986 ± 0.023012	0.960192 ± 0.034854
13 th	0.697512 ± 0.025461	0.890253 ± 0.022778	0.201406 ± 0.030348	0.960534 ± 0.033730
14 th	0.703224 ± 0.019900	0.902898 ± 0.019830	0.205575 ± 0.038635	0.932755 ± 0.033616
15 th	0.692050 ± 0.027536	0.891654 ± 0.018239	0.200860 ± 0.024085	0.929006 ± 0.031675
16 th	0.654609 ± 0.028141	0.928181 ± 0.013477	0.202355 ± 0.017180	0.963143 ± 0.033232
17 th	0.683020 ± 0.025644	0.890392 ± 0.023586	0.198464 ± 0.033714	0.941025 ± 0.034951
18 th	0.699170 ± 0.025487	0.911382 ± 0.027290	0.188802 ± 0.036537	0.950874 ± 0.028669
19 th	0.655108 ± 0.031857	0.884214 ± 0.027816	0.209234 ± 0.032466	0.962966 ± 0.034187
20 th	0.637446 ± 0.032562	0.873755 ± 0.036792	0.196807 ± 0.026001	0.952410 ± 0.038702

TABLE 1 : Résultats numériques des simulations pour chaque arrondissement avec $N = 50$ répétitions. Chaque valeur des indicateurs factice est donnée par sa moyenne sur les répétitions et la déviation standard associée. Le ratio de robustesse est calculé par rapport au premier arrondissement (choix arbitraire). Un ratio inférieur à 1 signifie que la borne de l'intégrale est plus petite pour le premier système, i.e. que l'évaluation est plus robuste pour celui-ci.

médian X_i . On définit des poids spatiaux utilisés pour quantifier l'intensité des interactions géographiques entre unités i, j , avec d_{ij} distance euclidienne entre centroïdes : $w_{ij} = \frac{P_i P_j}{(\sum_k P_k)^2} \cdot \frac{1}{d_{ij}}$ si $i \neq j$ et $w_{ii} = 0$. Les indicateurs normalisés sont les suivants

- Indice d'autocorrelation spatiale de Moran, défini comme la covariance pondérée normalisée du revenu médian par $\rho = \frac{N}{\sum_{ij} w_{ij}} \cdot \frac{\sum_{ij} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$
- Indice de dissimilarité (proche du Moran mais intégrant les dissimilarités locales plutôt que les corrélations), donné par $d = \frac{1}{\sum_{ij} w_{ij}} \sum_{ij} w_{ij} |\tilde{X}_i - \tilde{X}_j|$
avec $\tilde{X}_i = \frac{X_i - \min(X_k)}{\max(X_k) - \min(X_k)}$
- Le complémentaire de l'entropie de la distribution des revenus, qui est une façon de capturer des inégalités globales $\varepsilon = 1 + \frac{1}{\log(N)} \sum_i \frac{X_i}{\sum_k X_k} \cdot \log \left(\frac{X_i}{\sum_k X_k} \right)$

De nombreuses mesures de ségrégation avec différentes signification à différentes échelles existent, comme par exemple à l'échelle d'une unité spatiale élémentaire par comparaison de la distribution de revenus empirique avec un modèle nul [166]. Le choix est ici arbitraire, afin d'illustrer la méthode avec un nombre raisonnable de dimensions.

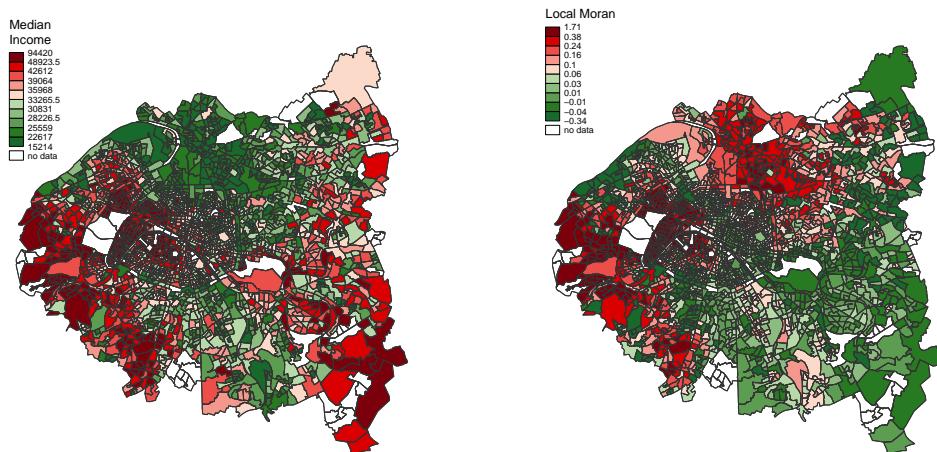


FIGURE 2 : Cartes de ségrégation métropolitaine. Les cartes montrent le revenu annuel médian pour les unités statistiques élémentaires (IRIS) pour les trois départements correspondant globalement à la métropole du Grand Paris, et l'index local d'autocorrelation spatiale de Moran correspondant, défini pour l'unité i par $\rho_i = N / \sum_j w_{ij} \cdot \frac{\sum_j w_{ij}(X_j - \bar{X})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}$. Les zones les plus ségréguées coincident avec les plus riches et les plus pauvres, suggérant une augmentation de la ségrégation dans les cas extrêmes.

RÉSULTATS La méthode est appliquée avec ces indicateurs à la zone du Grand Paris, constitué de 4 département qui sont des niveaux

administratifs intermédiaires. La création récente d'un nouveau système de gouvernance métropolitaine [109] met en évidence des interrogations sur sa pertinence, notamment sur ses capacités d'atténuer les inégalités spatiales. On peut voir en Fig. ?? les cartes de la distribution spatiale du revenu médian et de l'index local d'autocorrelation spatiale correspondant. La dichotomie bien connue entre est et ouest est retrouvée ainsi que la disparité des quartiers intra-muros, comme cela été présenté par diverses études, comme [118] à travers l'analyse des dynamiques des transactions immobilières. Notre cadre d'étude est ensuite appliqué à une question concrète ayant des implications pour la prise de décision : *dans quelle mesure une évaluation de la ségrégation au sein de différents territoires est sensible aux données manquantes*? Pour cela, on procède à des simulations de Monte-Carlo (75 répétitions) pour lesquelles une proportion fixe de données est supprimée aléatoirement, et l'indice de robustesse correspondant est évalué avec les indicateurs normalisés. Les simulations sont faites sur chaque département de façon indépendante, à chaque fois pour une robustesse relative à l'évaluation du Grand Paris complet. Les résultats sont présentés en Fig. ???. Toutes les zones ont une robustesse légèrement meilleure que la référence, ce qui pourrait être expliqué par une homogénéité locale et donc des indices de ségrégation plus fiables. Les implications pour la prise de décision qui peuvent être par exemple tirées sont des comparaisons directes entre les zones : une perte de 30% de l'information sur le 93 correspond à une perte de seulement 25% pour le 92. La première zone étant déjà défavorisée socio-économiquement, l'inégalité est augmentée par cette qualité moindre de l'information statistique. L'étude des déviations standard suggère des études plus approfondies comme différents régimes de réponse à la suppression de données semblent exister.

2.8.4 Discussion

Applicabilité à des situations réelles

IMPLICATIONS POUR LA PRISE DE DÉCISION L'application de notre méthode à des situations concrètes de prise de décision peu être pensée de différentes manières. Tout d'abord dans le cas d'un processus multi-attributs à but comparatif, comme la détermination d'un corridor pour une nouvelle infrastructure de transport, l'identification des territoires sur lesquels l'évaluation pourrait être biaisée (i.e. avec une mauvaise robustesse relative) devrait permettre une attention particulière pour ceux-ci, et l'adaptation des jeux de données ou la révision des points en conséquence. Dans tous les cas le processus total devrait être plus fiable. Une autre possibilité ressemble à l'application réelle que nous avons développé, i.e. la sensibilité de l'évaluation à divers paramètres comme les données manquantes. Si une décision paraît fiable car la taille de données est grande, mais

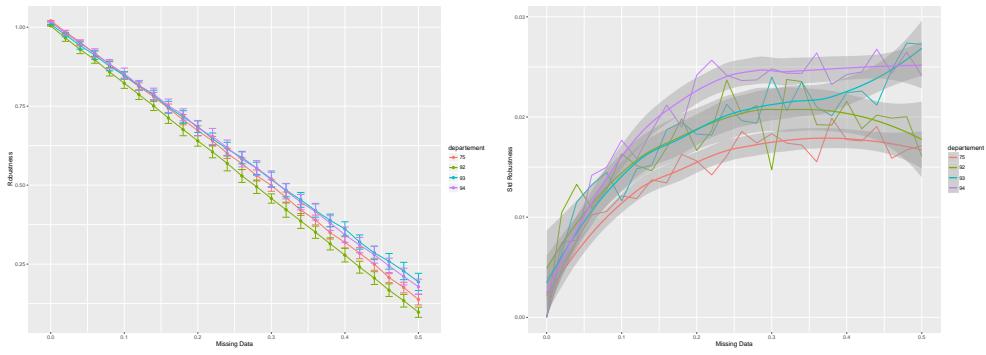


FIGURE 3 : Sensibilité de la robustesse aux données manquantes. Gauche.

Pour chaque département, des simulations de Monte-Carlo ($N=75$ répétitions) sont utilisées pour déterminer l'impact des données manquantes sur la robustesse de l'évaluation de la ségrégation. Les ratios de robustesse sont tous calculés relativement à la région métropolitaine complète avec toutes les données disponibles. Le comportement quasi-linéaire traduit une décroissance approximativement linéaire de la discrépance en fonction de la taille des données. Les trajectoires similaires des départements les plus pauvres (93,94) suggère que la correction au comportement linéaire est fonction des motifs de ségrégation. *Droite.* Déviations standard des ratios de robustesse. Les différents régimes (en particulier le 93 contre les autres) révèlent des transitions de phase à différents niveaux de données manquantes, signifiant que l'évaluation dans le 94 est de ce point de vue plus sensible aux données manquantes.

que l'évaluation est très sensible à la suppression de données, il faudra être prudent pour l'interprétation des résultats et pour la prise de décision finale. Un travail approfondi et de test sera cependant nécessaire pour comprendre le comportement du cadre dans différents contextes et pouvoir piloter son application dans des situations réelles diverses.

INTÉGRATION AU SEIN DE CADRES EXISTANTS L'applicabilité de la méthode à des cas réels dépendra directement de son intégration potentielle dans des environnements existants. Au delà des difficultés techniques qui apparaissent nécessairement en essayant de coupler ou d'intégrer des implémentations existantes, des obstacles plus théoriques pourraient émerger, comme des formulations floues des fonctions ou des types de données, la cohérence des bases de données, etc. De tels cadres multi-critères sont nombreux. Un développement possible serait l'intégration dans un cadre open-source, comme par exemple celui décrit dans [244] qui calcule divers indices de ségrégation urbaine, comme on l'a déjà illustré pour l'application à la ségrégation métropolitaine.

DISPONIBILITÉ DES DONNÉES BRUTES De manière générale, des données sensibles comme des questionnaires de transport, ou des données de sondage à granularité très fine, ne sont pas disponibles de manière ouverte, mais fournis de manière déjà agrégée à un certain niveau (comme par exemple les données françaises de l'Insee sont disponibles publiquement au niveau des unités statistiques élémentaires ou pour des zones plus grandes selon les variables et des contraintes de population minimale, les données plus précises étant à accès restreint). Cela signifie que l'application de notre cadre peut impliquer une procédure de recherche de données laborieuse, l'avantage d'être flexible étant alors compensé par ces contraintes additionnelles.

Validité des hypothèses théoriques

Une limitation possible de notre approche est la validité de l'hypothèse qui formule les indicateurs comme des intégrales spatiales. En fait, de nombreux indicateurs socio-économiques ne dépendent pas nécessairement directement de l'espace, et essayer de les associer à des coordonnées peut entraîner sur une pente glissante (par exemple, associer des variables économiques individuelles à des coordonnées résidentielles aura un sens seulement si la variable à une relation à l'espace, autrement un devient un artefact superflu). Même des indicateurs qui ont une valeur spatiale peuvent dériver de variables non-spatiales, comme [146] le souligne au sujet de l'accessibilité, en opposant les mesures d'accessibilité intégrée aux mesures individu-centrées mais pas forcément basée sur l'espace (comme par exemple des décisions individuelles). Contraindre une représentation

théorique d'un système pour le faire rentrer dans un cadre en changeant certaines de ses propriétés ontologiques (toujours dans le sens de la signification réelle des objets) peut être compris comme une violation d'une des règles pour la modélisation et la simulation en sciences sociales données par [14], car cela impliquerait qu'il pourrait exister un langage universel pour la modélisation, malgré qu'il ne puisse retranscrire certains systèmes, ayant pour conséquences des conclusions errantes à cause d'une rupture d'ontologie dans le cas d'une formulation sur-contrainte.

Généralité du Cadre

Nous soutenons qu'un des avantages fondamentaux de notre cadre est sa généralité et sa flexibilité, puisque la robustesse des évaluations est obtenue seulement par la structure des données si l'on relaxe les hypothèses sur les valeurs des poids. Des approfondissement pourraient inclure une formulation plus générale, en supprimant par exemple l'hypothèse d'agrégation linéaire. Des fonctions d'agrégation non-linéaires demanderaient toutefois de vérifier certaines propriétés regardant les inégalités intégrales. Par exemple, des résultats similaires pourraient être obtenus en s'orientant vers des inégalités intégrales pour fonctions Lipschitziennes, comme les résultats en une dimension de [93].

Conclusion

Nous avons proposé un cadre indépendant du modèle pour comparer la robustesse d'évaluations multi-attributs entre différents systèmes urbains. A partir de la discrépance des données, on fournit une définition générale de la robustesse relative sans aucune hypothèse de modèle pour le système, mais en supposant une agrégation linéaire des objectifs et des indicateurs exprimés comme des intégrales à noyaux. Nous proposons une première implémentation preuve de concept pour la ville de Paris pour laquelle les résultats numériques confirment la tendance générale attendue, et une implémentation sur des données réelles pour la ségrégation de revenus pour la région métropolitaine du Grand Paris, fournissant des réponses possibles à des questions de prise de décision plus concrètes. Des développements possibles peuvent inclure une analyse de sensibilité de la méthode, des applications à d'autres cas réels et une relaxation des hypothèses théoriques, c'est à dire de l'agrégation linéaire et de l'intégration spatiale.

3

QUANTITATIVE EPISTEMOLOGY

The Social Construction of What ?

- IAN HACKING [[121](#)]

Under this provocative book title by HACKING are implied complex mechanisms in the production of scientific knowledge. Animated debates on constructivism would be due to different metaphysical conceptions that are by essence not provable. As we have already evoked with perspectivism, scientific enterprises may have different purposes and be difficultly transferable to other contexts as we intent to do in our broad thematic vision developed in chapter [1](#).

A corollary of theoretical background proposed in chapter [8](#) is the need of an understanding of involved disciplines themselves to be able to build integrated heterogeneous models. The potentialities of couplings and integrations are greatly determined by existing approaches and corresponding gaps. This implies an advanced epistemological study in each field, that we propose to tackle in a systematic and quantitative way. This deliberate choice may shadow elaborated epistemological considerations but fits our purpose of preliminary investigations for the construction of models, as it may reveal investigation directions.

We describe and explore in a first section a systematic review exploration algorithm, that retrieve corpuses of references through iterative semantic extraction. We describe then briefly possible extended bibliometrics by presenting an external example of application. We finally suggest possible development directions towards unsupervised data and text-mining.

3.1 REVUE SYSTÉMATIQUE ALGORITHMIQUE

Une étude bibliographique étendue suggère une rareté des modèles quantitatifs de simulation qui intègrent à la fois la croissance urbaine et la croissance des réseaux. Cette absence pourrait être due aux intérêts divergents des disciplines concernées qui induiraient un manque de communication. Nous proposons de procéder à une revue de la littérature systématique et algorithmique pour donner des éléments de réponse quantitatifs à cette question. Un algorithme itératif formel pour construire des corpus de références à partir de mots-clés initiaux, basé sur l'analyse textuelle, est développé et mis en oeuvre. Nous étudions ses propriétés de convergence et procédons à une analyse de sensibilité. Nous l'appliquons ensuite à des requêtes représentatives de notre question spécifique, pour lesquelles les résultats tendent à confirmer l'hypothèse d'isolation des disciplines.

3.1.1 *En recherche de modèles de co-évolution*

Les réseaux de transport et l'usage du sol urbain sont connus pour être des composantes fortement couplées des systèmes urbains à différentes échelles [50]. Une approche commune est de les considérer comme étant en co-évolution, tout en évitant les interprétations trompeuses comme le mythe des effets structurants des infrastructures de transport [190]. Une question qui se présente rapidement est l'existence de modèles endogénéisant cette co-évolution, i.e. prenant en compte simultanément la croissance urbaine et celle du réseau. Nous essayons d'y répondre par une revue systématique algorithmique. Nous proposons dans cette section, après un état de l'art rapide de la littérature existante, de développer cette approche en formalisant l'algorithme, dont les résultats sont ensuite présentés et discutés.

3.1.2 *Modéliser les Interactions entre croissance urbaine et croissance des réseaux : une revue*

Modèles LUTI

Une large classe de modèle développés essentiellement dans des objectifs de planification, les modèles d'interaction entre transport de usage du sol, sont un premier type pouvant rentrer dans notre problématique. Voir les diverses revues [57], [136] et [259] pour avoir un aperçu de l'hétérogénéité des approches incluses, qui existent depuis plus de 30 ans. Des versions récentes avec divers raffinements sont toujours développés aujourd'hui, comme [86] qui inclut le marché immobilier pour la région parisienne. Différents aspects du même système peuvent être traduits par divers modèles (comme e.g. [260]), et le trafic, les dynamiques résidentielles et d'emploi, l'évolution de

l'usage du sol en découlant, influencée aussi par un réseau de transport statique, sont généralement pris en compte.

Approches de Croissance de Réseau

A l'opposé de nombreux travaux ont pris la logique inverse, i.e. essayent de reproduire la croissance du réseau étant donné des hypothèses sur l'environnement urbain, comme résumé dans [276]. Dans [267], les travaux économiques empiriques sont positionnés parmi les autres approches de la croissance des réseaux, comme des travaux de physiciens proposant des modèles de croissance géométrique locale [19]. L'analogie avec la biologie a également déjà été faite, permettant de reproduire les propriétés typiques de robustesse des réseaux de transport [243].

Approches hybrides

Peu de travaux couplant croissance urbaine et croissance du réseau sont disponibles dans la littérature. [20] couple l'évolution de la densité avec la croissance du réseau dans un modèle jouet. Dans [217], un automate cellulaire simple couplé à un réseau évolutif reproduit les faits stylisés des Etablissements Humains décrits par Le Corbusier. A une plus petite échelle, [2] propose un modèle de co-évolution entre routes et bâtiments, en suivant des règles géométriques. Ces approches restent cependant limitées et rares.

3.1.3 Analyse Bibliométrique

La revue de littérature est une étape préliminaire cruciale à toute entreprise scientifique, et sa qualité et son étendue peut avoir un impact conséquent sur le résultat final. Des techniques de revue systématique ont été développées, des revues qualitatives aux meta-analyses quantitatives qui permettent de produire des nouveaux résultats par combinaison d'études existantes [226]. Passer sous silence certaines références peut même être considéré comme une erreur scientifique dans le contexte de l'émergence des systèmes d'information [160]. Nous proposons de tirer parti de telles techniques pour traiter notre problème. En effet, l'observation de la bibliographie obtenue dans la section précédente soulève une hypothèse. Il semble clair que toutes les briques sont présentes pour l'existence de modèles co-évolutifs mais des questionnements et objectifs différents semblent la stopper. Comme montré par [68] pour le concept de mobilité, pour lequel un "petit monde d'acteurs" relativement fermé a inventé une notion ad hoc, utilisant des modèles sans connaissance préalable d'un contexte scientifique plus général, on pourrait se trouver dans un cas similaire pour le type de modèles auxquels on s'intéresse. Des interactions restreintes entre des champs scientifiques travaillant sur

les mêmes objets mais avec des objectifs et contextes divergents, et à des échelles différentes, pourrait être à l'origine de l'absence de modèles co-évolutifs. Tandis que la majorité des études en bibliométrie se reposent sur les réseaux de citation [184] ou les réseaux de co-auteurs [231], nous proposons d'utiliser un paradigme moins exploré, basé sur l'analyse textuelle, introduit par [59], qui obtient une cartographie dynamique des disciplines scientifiques en se basant sur leur contenu sémantique. La méthode est particulièrement adaptée pour notre étude puisque nous voulons comprendre la structure du contenu des recherches sur le sujet. Nous appliquons une approche algorithmique décrite par la suite. L'algorithme procède par itérations pour obtenir un corpus stabilisé à partir de mots-clés initiaux, reconstruisant l'horizon sémantique scientifique autour d'un sujet donné.

Description de l'Algorithme

Soit \mathcal{A} un alphabet, \mathcal{A}^* les mots correspondants et $T = \cup_{k \in \mathbb{N}} \mathcal{A}^{*k}$ les textes de longueur finie sur celui-ci. Ce qu'on nomme une référence est pour l'algorithme un enregistrement avec des champs textuels représentant le titre, le résumé et les mots-clés. L'ensemble de références à l'itération n sera noté $\mathcal{C} \subset T^3$. Nous supposons l'existence d'un ensemble de mots-clés \mathcal{K}_n , les mots-clés initiaux étant \mathcal{K}_0 . Une itération procède de la manière suivante :

1. Un corpus intermédiaire brut \mathcal{R}_n est obtenu par une requête à un catalogue auquel on fournit les mots-clés précédents \mathcal{K}_{n-1} .
2. Le corpus total est actualisé par $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \mathcal{R}_n$.
3. Les nouveaux mot-clés \mathcal{K}_n sont extraits du corpus par Traitement du Langage Naturel (NLP), étant donné un paramètre N_k fixant le nombre de mot-clés.

L'algorithme termine quand la taille du corpus devient stable ou quand un nombre maximal d'itérations défini par l'utilisateur est atteint. La figure 4 montre le processus général.

Résultats

IMPLÉMENTATION De par l'hétérogénéité des opérations requises par l'algorithme (organisation des références, requêtes au catalogue, analyse textuelle), le langage Java s'est présenté comme une alternative raisonnable. Le code source est disponible sur le dépôt ouvert du projet¹. Les requêtes au catalogue, qui consistent à récupérer un ensemble de références à partir d'un ensemble de mots-clés, sont faites via l'API du logiciel Mendeley [180] qui permet un accès ouvert à une base de données conséquente. L'extraction des mots-clés est effectuée

¹ à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR>

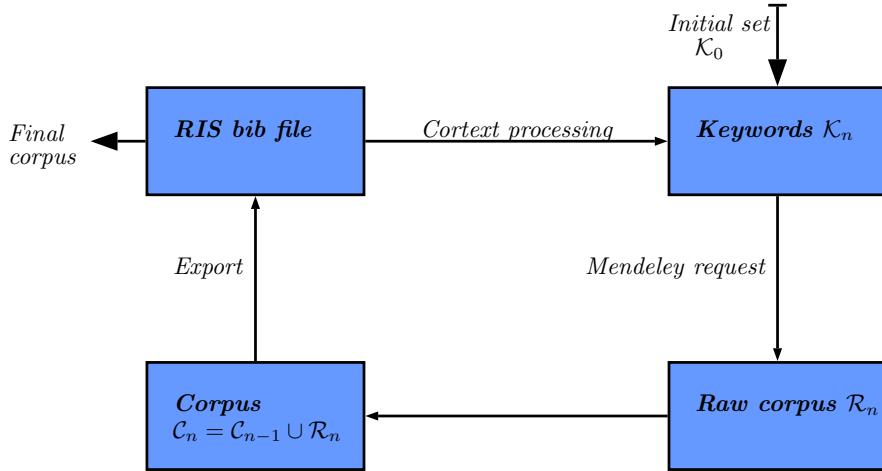


FIGURE 4 :

par techniques d’Analyse Textuelle (NLP) selon le processus donné dans [59], via un script Python qui utilise [34].

CONVERGENCE ET ANALYSE DE SENSIBILITÉ Une preuve formelle de convergence de l’algorithme n’est guère envisageable puisque qu’elle dépendra de la structure empirique inconnue des résultats de requête et d’extraction de mots-clés. Il est donc nécessaire d’étudier le comportement de l’algorithme de manière empirique. Comme présenté en figure 5, l’algorithme a de bonnes propriétés de convergence mais diverse sensibilités à N_k . Nous étudions également la cohérence lexicale interne des corpus finaux et fonction du nombre de mots-clés. Comme attendu, des valeurs faibles produisent des corpus plus cohérents, mais la variabilité lorsque qu’elles augmentent reste raisonnable.

Lorsque l’algorithme a été partiellement validé, on peut l’appliquer à notre question. Nous partons de cinq différentes requêtes initiales qui ont été manuellement extraites des divers domaines identifiés dans la bibliographie (qui sont “city system network”, “land use transport interaction”, “network urban modeling”, “population density transport”, “transportation network urban growth”). Nous prenons l’hypothèse la plus faible pour le paramètre $N_k = 100$, au sens où les domaines atteints devraient être moins restreints. Après avoir construit les corpus, nous étudions leur cohérence lexicale comme un indicateur de réponse à notre question initiale. De grande distances devraient confirmer l’hypothèse formulée ci-dessus, i.e. que des disciplines auto-centrées pourraient être à l’origine d’un manque d’intérêt pour des modèles co-évolutifs. La table 2 montre les valeurs de la proximité lexicale relative, qui est significativement basse, confirmant notre hypothèse.

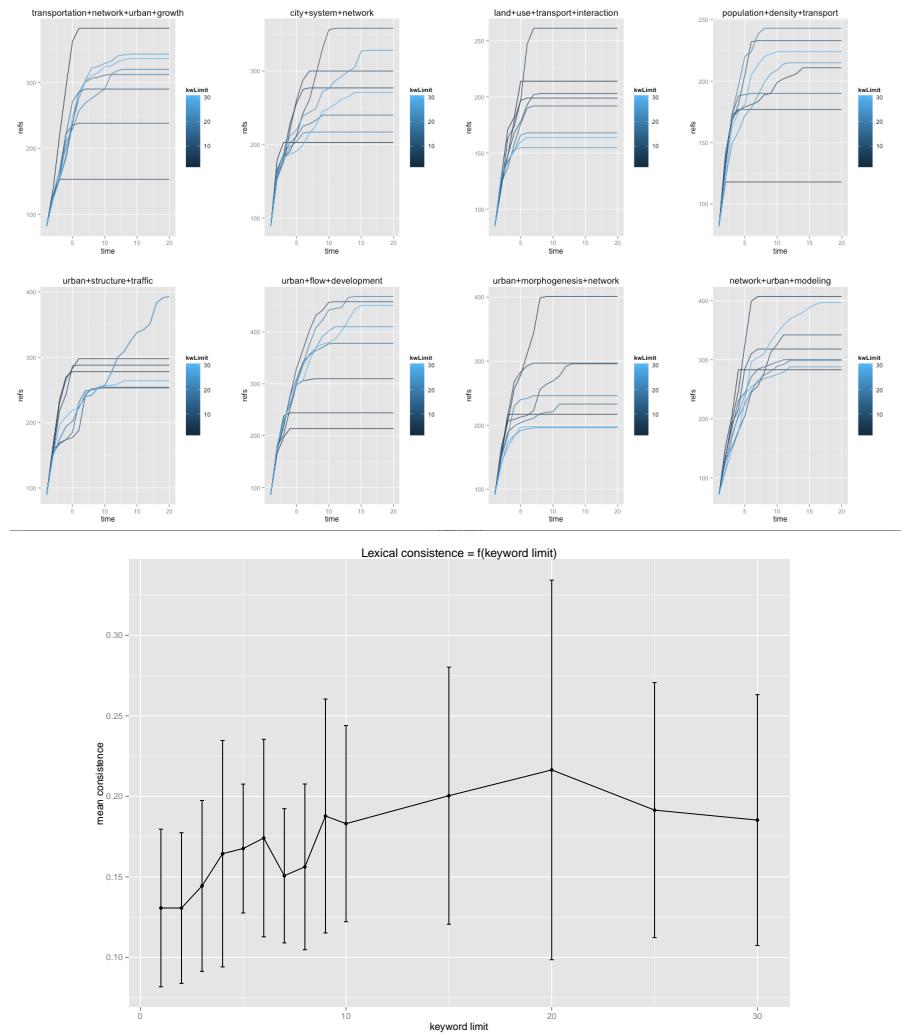


FIGURE 5 :

Corpus	1	2	3	4	5
1 (W=3789)	1	0	0.0719	0.0078	0.0724
2 (W=5180)	0	1	0.0338	0	0.0125
3 (W=3757)	0.0719	0.0338	1	0.0100	0.1729
4 (W=3551)	0.0078	0	0.0100	1	0.0333
5 (W=8338)	0.0724	0.0125	0.1729	0.0333	1

TABLE 2 :

Les développements possibles incluent la construction de réseaux de citation via un accès automatique à Google Scholar qui fournit les citations entrantes. La confrontation des coefficients inter-clusters pour le réseau de citations entre les différents corpus avec la cohérence lexicale est un aspect clé d'une validation approfondie des résultats.

L'absence peu explicable a priori de modèles qui simulent la coévolution des réseaux de transport et de l'usage du sol urbain, qui se confirme à première vue par un état de l'art couvrant des domaines disparates, pourrait être due à l'absence de communication entre les disciplines scientifiques étudiant différents aspects du problème. Nous avons proposé une méthode algorithmique pour donner des éléments de réponse par l'extraction de corpus basée sur l'analyse textuelle. Les premiers résultats numériques semblent confirmer l'hypothèse. Cependant, une telle analyse quantitative ne doit pas être considérée seule, mais devrait plutôt venir comme soutien à des études qualitatives qui peuvent être l'objet de développements futurs, comme celle menée dans [68], dans laquelle des questionnaires avec des acteurs historiques fournissent des informations extrêmement pertinentes.

3.2 BIBLIOMÉTRIE INDIRECTE PAR ANALYSE DE RÉSEAUX COMPLEXES

As described before, semantic analysis does not contain all the information on disciplinary compartmentation nor on patterns of propagation of scientific knowledge as the ones contained in citation networks for example. Furthermore, data collection in the previous algorithm is subject to convergence towards self-consistent themes because of the proper structure of the method. It may give more insight about scientific social patterns of ontological choices in modeling to study communities in broader networks, that would more correspond to disciplines (or sub-disciplines depending on granularity level).

TODO : insert / adapt / translate cybergeo paper : run on a specific corpus

3.2.1 Application Spécifique

We will try to reconstruct the same way disciplines around our thematic, and by for example identifying bridge articles (nodes with high centrality or vulnerability) identify crucial thematic elements and research directions.

An other application will be the reflexivity of our thesis : we attend to proceed to similar analysis on our proper bibliography (and its evolution, available via git history), to understand our patterns of knowledge, possible gaps or unveil unexpected developments.

3.3 VERS UNE MODÉLISATION DES THÈMES ET UNE EXTRACTION AUTOMATIQUE DU CONTEXTE

A possible direction to strengthen our quantitative epistemological analysis would be to work on full texts related to the modeling of interaction between networks and territories, with the aim to automatically extract thematics within articles. The idea would be to perform some kind of automatized modelography, with possible features to be extracted that would be ontologies, model architecture or structures, scales, or even typical parameter values. It is not clear to what degree structure of models can be extracted from their description in papers and it surely depends on the discipline considered. For example in a framed field such as transportation planning, using a pre-defined ontology (in the sense of dictionary) and a fuzzy grammar could be efficient to extract information as the discipline is relatively formatted. In theoretical and quantitative geography, beyond the barrier of language, information organisation is surely less subject to unsupervised data-mining because of the more literary nature of the discipline : synonyms and figures of speech are generally the norm in good level human sciences writing, fuzzing a possible generic structure of knowledge description.

Depending on extended results of the two previous sections and on thematic requirements (huge need of knowledge on precise models structure, that may appear when trying to construct more specialized operational models), this project may be conducted with more or less investment.

Deuxième partie

MATERIALS

This part aims at producing knowledge from the empirical analysis of case studies and from first modeling experiments. Explicit testing of hypothesis drawn from the theory is not achieved yet as these are preliminary steps for a reasoned insight into empirical and modeling domains.

4

EMPIRICAL ANALYSIS : INSIGHTS FROM STYLIZED FACTS

*Mais ce n'est pas une question
d'âge, de chiffres et de stats
Moi je te parle surtout de rage, de kif
et d'espoir*

- YOUSSEOPHA , Esperance de Vie

As this quote suggests, a purely quantitative view of the world makes no sense without qualitative counterbalancing. More precisely, we argue that the *cliché* of an opposition between quantitative and qualitative analysis is an illusion. No distinct boundary exists between both. We propose to call quantitative any process involving computation by a Turing machine, whereas the qualitative will be for us the modeling design process and its interpretations. Therefore both are necessarily closely interlaced in any of our approaches. In particular concerning the construction and the validation or refutation of our theory, empirical analysis on real case studies, implying the extraction and qualification of stylized facts, follows that schema.

We propose in this chapter various empirical analysis on different objects at different scales. A first section begins the examination of static spatial correlations between morphological measures of population density and road network measures on Europe at a 500m resolution. Applying last section of the methodological chapter should provide information on typical spatial scales of interaction between these indicators of territory and network and on dynamical correlations between these. These computation furthermore provide empirical measures on which one model will be calibrated. We then describe a roadmap for statistical analysis on dynamical data of interactions for Bassin Parisien in the last fifty years. An other project using Real Estate transaction data for Parisian Metropolitan Region aim at seeking early warning of network breakdowns. We finally describe potential analyses on South African historical data.

4.1 CORRÉLATIONS STATIQUES ENTRE FORME URBAINE ET FORME DE RÉSEAU

Spatio-temporal processes implying diffusion or propagation phenomena generally have a specific structure of correlation. In particular, as derived in section 2.5, a static computation of correlation between different instances of a system may under certain conditions provide information on dynamical correlations implied.

4.1.1 Mesures morphologiques de la densité de population européenne

Contexte

A l'échelle macroscopique du système de ville, le caractère spatial du système urbain est capturé de manière raisonnable par les positions des villes, associées aux variables agrégées au niveau de la ville qui représentent entièrement le système (voir e.g. l'ontologie des modèles Simpop [206] ou de leur successeur Marius [70]). A l'échelle mesoscopique, à laquelle nous nous attendons à capturer des manifestations morphologiques des interactions entre ville et transport, la structure du système territorial peut être spécifiée par des indicateurs plus raffinés pour l'aspect morphologique.

Analyse Empirique

We study systematically morphological indicators for constant size areas covering European Community. The choice of fixed size areas can be questioned regarding definition of a territorial system, that can be otherwise understood as a consistent spatial entity at a given scale and along certain criteria : *Human territories* as defined by Raffestin (op. cit.) or more generally functionally autonomous spaces¹. Here we choose the mesoscopic scale of a metropolitan center ($\simeq 50\text{km}$) for comparability purposes and because greater scale are no more relevant regarding urban form, whereas smaller scales must contain too much noise.

Data is the European population density grid [99] and indicators computation is implemented in parallel using R with Fast convolution raster functions. We show in next figures computed values of morphological indicators (see [153] for a precise formulation of indicators that are Moran index, average distance, entropy and hierarchy).

¹ for example, a tentative of definition of a *Parisian* territory would present many facets. From the subjective territory point of view, intra-muros Parisians consider a strict boundary at *Boulevard Peripherique*, whereas close and even further suburbs will be seen as Parisians from the Province. The functional territory of *Metropolitain* extends slightly further than the administrative boundary. Governance perimeters are currently mutating with the Metropolitan governance project. Complementary perceptions of the territory can thus be multiplied.

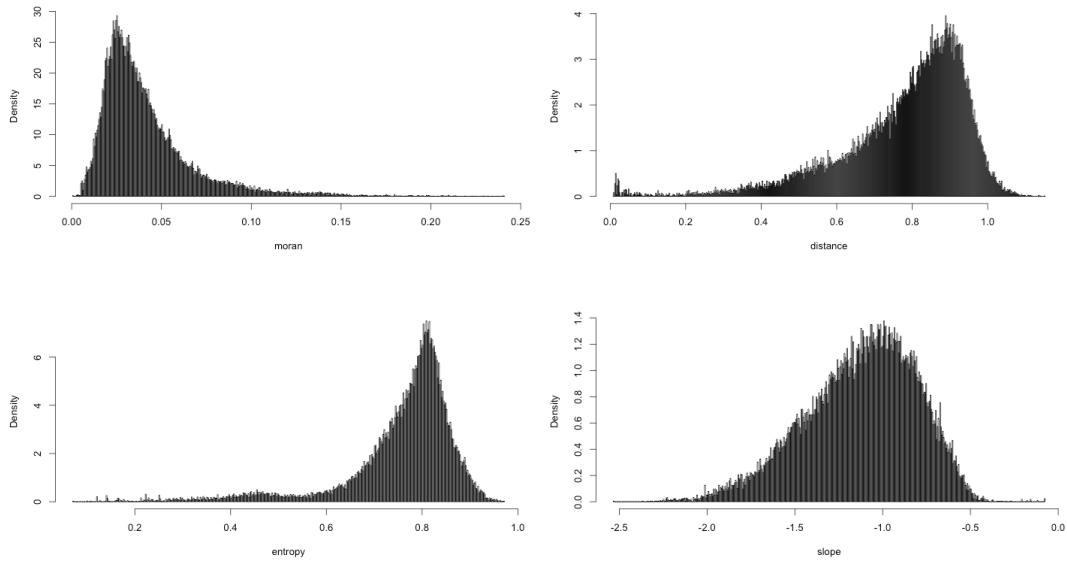


FIGURE 6 : Distribution empirique des indicateurs morphologiques

Développements

In [239] density grids for other countries across the world (ex. China) are provided² so we may repeat our analysis to other regions for comparison purposes.

4.1.2 Mesures de Réseau

We consider network aggregated indicators as a way to characterize transportation network properties on a given territory, the same way morphological indicators yielded information on urban structure. We propose to compute some simple indicators on same extents as for morphology, to be able to explore relations between these static measures. Static network analysis has been extensively documented in the literature, see [164] for a cross-sectional study of cities or [150] for exploration of new measures for the road network.

Pré-traitement des données

We work in a first time on road network, which structure is finely conditioned to territorial configuration of population densities. Furthermore, data for present day road network is available through the OpenStreetMap project [192]. Its quality was investigated for different countries such as England [124] and France [110]. It was found to be of a quality equivalent to official surveys for the primary road network.

² available at <http://www.worldpop.org.uk/>

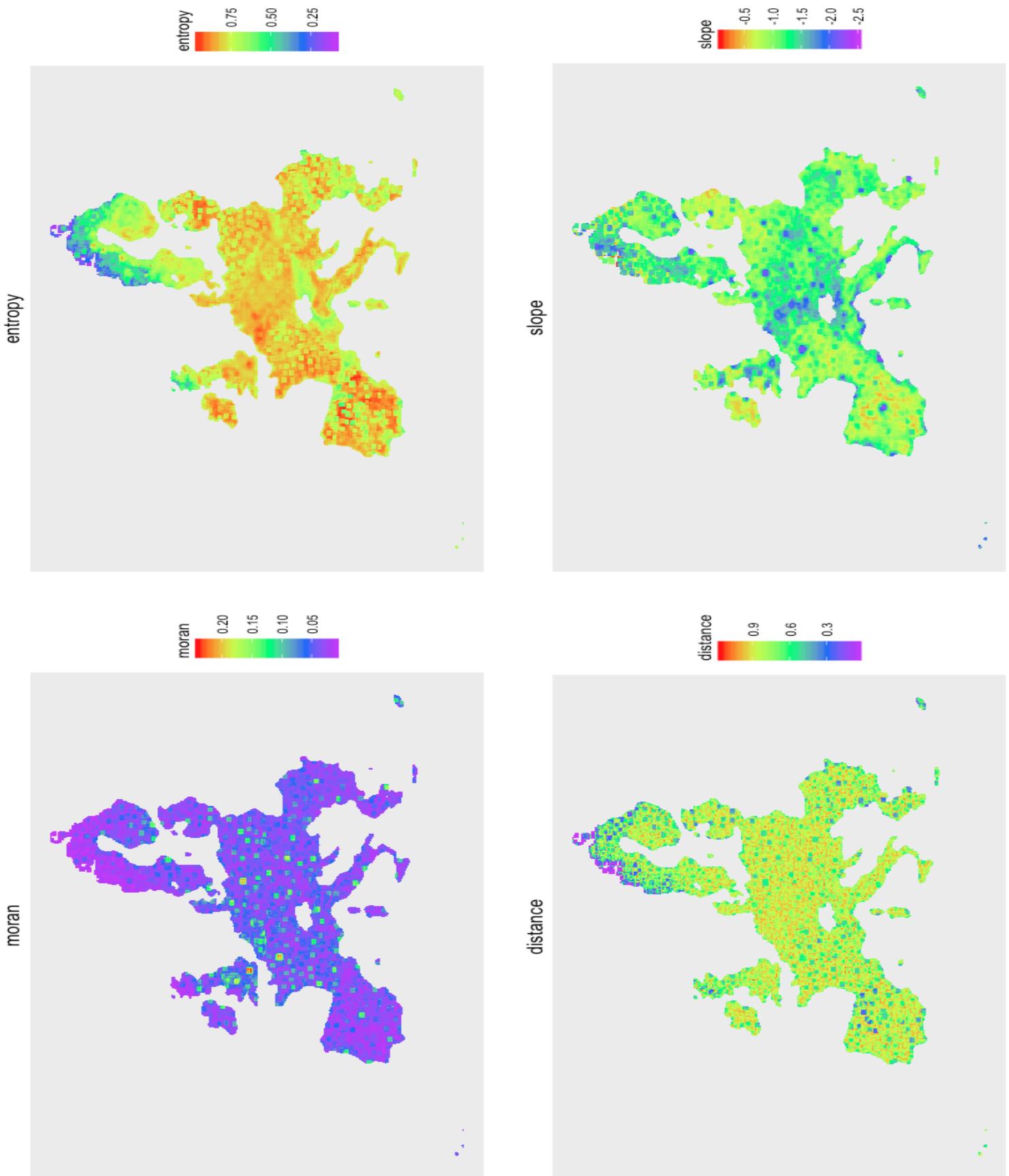


FIGURE 7 :

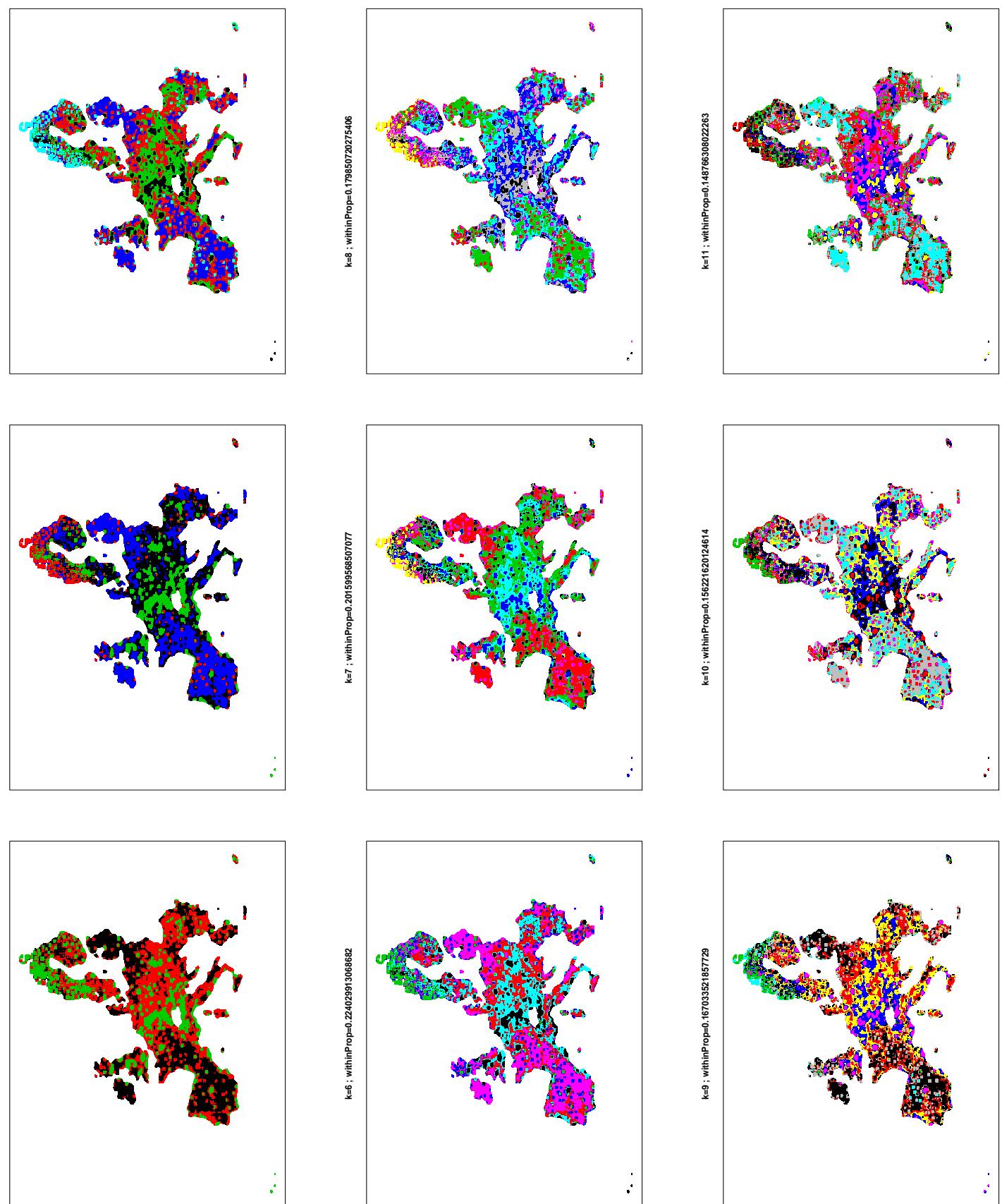


FIGURE 8 :

ALGORITHME DE SIMPLIFICATION For a given dataset corresponding to a subset of the overall road network, it is necessary to simplify network structure by spatial aggregation as initial data presents very detailed features and thus a very large numbers of nodes ($\simeq 10^{10}$ for Europe dataset). Such a level of precision is not needed in our study since density data is already aggregated at 500m resolution. It is possible to drastically reduce network size by spatial aggregation of nodes and link replacements. More precisely we use the following procedure :

- a background raster (which resolution r gives the snapping parameter for aggregation) is constructed from a reference raster and the extent of network. This grid gives spatial aggregation units for network nodes.
- for each feature of the road dataset, corresponding connected raster cells are stored with corresponding impedance and distance in a sparse adjacency matrix.
- Network is simplified by iterative suppression of nodes with degree two, with keeping link speed and real length to their effective value.

IMPLÉMENTATION A PostGIS database is used to store raw and simplified network, in order to perform efficient spatial requests, compared for example to initial osm data formats (osm or pbf). However the size of storage of data into this base is much higher (factor 10) so processing was parallelized between european countries. Consistence is ensured by the use of the same common density raster as simplification canvas. Final network is stored into the Postgis database for efficient indicator computation given a spatial extent.

SENSIBILITÉ AUX PARAMÈTRES DE SIMPLIFICATION Sensitivity of indicators to raster resolution and to degree simplification algorithm must still be tested to ensure the relevance of data preprocesing.

Indicateurs

Network macroscopic structure is summarized by the following set of indicators, after the simplifications and reductions done in the previous step. Assuming network given by $N = (V, E)$, nodes spatial positions $\vec{x}(V)$ and edges *effective distances* $d(E)$ taking into account impedances and real distances (to include basically network hierarchy), we have indicators :

- connectivity
- degree distribution

- centrality, taken as normalized mean *betweenness-centrality*
- average path length
- network diameter
- mean network speed

These indicators are used to capture a rough picture of the structure. Refined work at smaller scales (intra-urban road network) and with more elaborated measures that allow to differentiate more precisely local form, was recently done by Lagesse in [150].

Résultats

Les indicateurs de réseau ont été calculés sur des zones similaires aux indicateurs de forme urbaine,

4.1.3 Correlations Statiques Effectives

4.1.4 Non-stationnarité spatiale et non-ergodicité

Case study : Context and Rationale

Study of interactions between network and territories :

→ searching for stylized facts, what can be learnt from static correlations between urban form and road network ?

Theoretical Background : A Theory of co-evolutive networked human territories proposed in [219], that in particular postulates an important role of networks in the morphogenesis of complex adaptive urban systems that are human territories

→ investigation of stationarity and ergodicity properties of relation between road network and population distribution ; implies spatiality of correlations and link static-dynamic

Dataset construction

Computation of topological road network for all Europe, at 100m granularity scale (to be used consistently with population grid [99])

→ Import of OSM into pgsql, simplification at 100m granularity, topological simplification with split/merge algorithm

$\simeq 44 \cdot 10^6$ links in initial OSM db, $\simeq 61 \cdot 10^6$ in first simplified layer,
 $\simeq 21 \cdot 10^6$ in final database

Results : Computation of Indicators

Computation of urban form indicators [153] and network indicators on $l_0 = 10\text{km side square}$

Results : Spatial Correlations

Computation of spatial correlation on square areas of width $\delta \cdot l_0$ (with typically $\delta = 4, \dots, 16$)

→ local spatial stationarity of processes

Results : Multi-scale Processes

→ Significant variation of mean correlation with δ (Left) and of normalized confidence interval (Right) given by $|\rho_+ - \rho_-| \cdot \delta$, as bounds theoretically vary as $\sqrt{N} \sim \sqrt{\delta^2}$: implies multi-scalarity

Empirical Findings (Formalization)

$Y_i[\vec{x}, t]$ spatio-temporal stochastic process, verifies empirically :

1. Local spatial autocorrelation is present and bounded by l_ρ (in other words the processes are continuous in space) : at any \vec{x} and t , $|\rho_{\|\Delta\vec{x}\| < l_\rho} [Y_i(\vec{x} + \Delta\vec{x}, t), Y_i(\vec{x}, t)]| > 0$.
2. Processes are locally parametrized : $Y_i = Y_i[\alpha_i]$, where $\alpha_i(\vec{x})$ varies with l_α , with $l_\alpha \gg l_\rho$ and weakly locally stationary in space.
3. Processes are multi-scalar : since $\rho(\delta = \infty) > \rho(\delta = 0)$, a necessary non-linear correction on processes spatial averages in correlation computation is present.

Analytical Deductions

1. **Regimes of temporal correlations.** Let assume local ergodicity in \vec{x}_0 at scale $\delta \cdot l_0$ (reasonable with urban growth and network extension in recent times). The Ergodic theorem implies that $\exists \mathcal{T}$ such that

$$\langle Y_i(t) \rangle_{\|\vec{x} - \vec{x}_0\| < \delta \cdot l_0} = \langle Y_i(\vec{x}_0) \rangle_{t \in \mathcal{T}}$$

With spatial stationarity, $\langle Y_i \rangle_{\vec{x}_0} = \langle Y_i \rangle_{\vec{x}_1}$, thus \mathcal{T} must be constant to be invariant by translation. By contraposition and (2), processes have different dynamical characteristics.

2. **Global non-ergodicity.** Let X_k a partition of space into local areas. We have $\langle \cdot \rangle_x = \sum_k w_k \langle \cdot \rangle_{x_k} = (1) \sum_k w_k \langle \cdot \rangle_{\mathcal{T}_k}$. On the other hand, global ergodicity would give $\langle \cdot \rangle_t = \langle \cdot \rangle_{\mathcal{T}} = \sum_k w_k \langle \cdot \rangle_{\mathcal{T}}$ and $\sum_k w_k (\langle \cdot \rangle_{\mathcal{T}} - \langle \cdot \rangle_{\mathcal{T}_k}) = 0$. Being true on each subset implies $\mathcal{T} = \mathcal{T}_k$, what contradicts (1).

Case study : implications

→ Still points to explore :

- variable correlations areas (size and shape in space)
- same work on cities population/train network data, which are also dynamical databases : extrapolation of ergodicity parameters ?

- correlations of returns : link between $\rho[\Delta_t Y]$ and $\rho[\Delta_x Y]$ (more difficult : if pure local ergodicity, \exists a permutation making the correspondance)
- Link between $\Delta_\delta \rho(\delta)$ and process derivatives ?

→ We show the regional nature of network-territories interactions, in particular the non-ergodicity of urban systems on **the interaction these components**

→ No direct results on time dynamics, but indirect : spatio-temporal processes do not have same speed and react/diffuse differently

4.1.5 Application

à

la

Chine

4.2 ISOLER LA CO-ÉVOLUTION DES RELATIONS CAUSALES

Spatial statistics studies on dynamical relations between network and territories are relatively rare. [157] does so on London metropolitan area and identifies causalities using lagged variables, but does not disentangle relations in the sense of coupled statistical models that would isolate endogenous effects from coupling effects.

4.2.1 Formalisation

We assume a dynamic transportation network $n(\vec{x}, t)$ within a dynamic territorial landscape $\vec{T}(\vec{x}, t)$, which components are to simplify population $p(\vec{x}, t)$ and employments $e(\vec{x}, t)$. Data is structured the following way :

- Observation of territorial variables are discretized in space and in time, i.e. the spatial field \vec{T} is summarized by $T = (\vec{T}(\vec{x}_i, t_j^{(T)}))_{i,j}$, with $1 \leq i \leq N$ and $1 \leq j \leq T$. They concretely correspond to census on administrative units (*communes* in our case) at different dates.
- Network has a continuous spatial position but is represented by the vector of network distances N

4.2.2 Sur

l'accessibilité

The notion of accessibility has been central to regional science since its introduction and systematization in planning around 1970. As already introduced in the first chapter, we question the notion of accessibility : *Is the notion of accessibility crucial for statistical analysis ?*

Weibull has proposed an axiomatic approach to accessibility [261], deriving a canonical decomposition for any *attraction-accessibility* function $A(a, d)$, assuming expected thematic axioms among others technical ones that are :

1. A is invariant regarding the order of the configuration
2. A decrease with distance at fixed attraction and increase with attraction at fixed distance
3. A is invariant when adding null attractions and constant configurations

Then A verifies these if and only if it is of the form

$$A[(a_i, d_i)] = T \left(\bigoplus_i z(d_i, a_i) \right)$$

where T is increasing with null origin, z is a *distance substitution function* (i.e. verifying axiom 2) and \oplus a *standard composition* associating two attractions at zero distance to the corresponding unique one.

It means that well suited matrices of autocorrelation should capture accessibility in regressions ; or it must be captured by non-linear regression on N . It may reveal some kind of intrinsic accessibility that is related to real phenomena (that we expect to fit with calibrated functions of accessibility based on Hedonic models e.g.) Seeing accessibility as a potential field is an equivalent vision : given any stationary dynamic for n, \vec{T} , Helmoltz theorem states that it derives from a potential (can be adapted to non-stationary dynamics with a time-varying potential).

4.2.3 Données

We will work on a novel dataset provided by LE NECHET, that consists in main road infrastructures with their opening dates and train network for network dynamics, and in population and employments of communes at census dates, for Bassin Parisien on the last fifty year. The temporal granularity due to census temporal step may be an obstacle to obtain good dynamical statistics.

4.2.4 Tests

Statistiques

The following large set of analysis are to be tested (non exhaustive) :

- On raw data :

- Multivariate models

$$\mathcal{L} [\mathbf{T}, \mathbf{N}] \sim \varepsilon$$

- Autocorrelated univariate models

$$(\mathbf{I} - \Sigma \mathbf{R} \mathbf{W}) \mathbf{X} \sim \varepsilon$$

- Autocorrelated multivariate models

$$(\mathcal{L}' - \Sigma \mathbf{R} \mathbf{W}) [\mathbf{T} + \mathbf{N}] \sim \varepsilon$$

- Geographically Weighted Regression [51]

$$\mathcal{L} [\mathcal{G} (\mathbf{T}, \mathbf{N})] \sim \varepsilon$$

- Granger causality tests : [266] use for example Granger causality to link transit with land-use changes.

- On data returns :

- Autoregressive multivariate models

$$\mathcal{L} [(\Delta T(t_{j'}))_{j' \leq j}, (\Delta N(t_{j'}))_{j' \leq j}] \sim \varepsilon$$

- Autoregressive autocorrelated multivariate models : idem with spatial autocorrelation term.
- Synthetic Instrumental Variables : static territory and/or network?

4.2.5 Méthode Générique

Description

Nous décrivons ici une méthode générique, basée sur un test similaire à la causalité de Granger [], pour tenter d'identifier des relations causales dans des systèmes spatiaux. Soit $X_j(\vec{x}, t)$ des processus aléatoires spatiaux unidimensionnels. Une réalisation d'un sous-système territorial est donnée par des ensembles de trajectoires pour chaque processus $x_{i,j,t}$. On suppose l'existence de fonctions de correspondance $\Phi_{j1,j2}$ permettant de faire correspondre les réalisations de chaque composantes à un index unique (dans le cas le plus simple, on associera les variables sur les mêmes patches). Si $\text{argmax}_{\tau} \hat{\rho}[x_{j1}, x_{j2}]$ est clairement défini, son signe donnera alors le sens de la causalité entre les composantes $j1$ et $j2$.

[168] : Spatio-temporal Granger causality for fMRI data ; not really spatial in the sense that no real distance effect.

Données Synthétiques

ETUDE DE CAS Cette méthode doit dans un premier temps être testée et partiellement validée, ce que nous proposons de faire sur des données synthétiques, approche dont l'utilisation est documentée et illustrée au chapitre ?? . [217] est un modèle simple de morphogenèse urbaine (modèle RBD) faisant un candidat intéressant pour notre test. En effet, les variables explicatives de la croissance urbaine, les processus d'extension du réseau et le couplage entre densité urbaine et réseau sont assez élémentaires. Cependant, hormis dans des cas extrêmes (distance au centre détermine valeur foncière uniquement, le réseau dépendra de manière causale de la densité, ou distance au réseau seule, la causalité devrait être inversée), les régimes mixtes n'exhibent pas de causalités évidentes : c'est donc un parfait cas pour tester si la méthode est capable d'en détecter.

APPLICATION Nous explorons une grille de l'espace des paramètres du modèle RBD. Pour chaque valeur des paramètres, nous procémons à $N =$ répétitions. Le modèle a par ailleurs été exploré de nouveau pour reproduction et extension des résultats.

4.3 TRAJECTOIRES DE MARCHÉS IMMOBILIERS

4.3.1 *Contexte*

Des aspects très variés des territoires sont concernés par l’interaction avec les réseaux. Dans nos études précédentes, aucun aspect socio-économique des populations habitant le territoire ni des valeurs économiques pour le foncier et l’immobilier n’ont été considérés. Il s’agit cependant d’éléments cruciaux des dynamiques territoriales et sont étudiés de manière intensive dans des champs comme l’analyse territoriale ou l’économie urbaine : par exemple, [134] étudie les choix résidentiels des ménages pour comprendre les interactions entre usage du sol et transport. Nous proposons ici d’utiliser une base de données de transactions immobilières pour la région parisienne sur les 20 dernières années, avec une granularité temporelle de 2 ans et coordonnées spatiales exactes. [118] l’utilise pour établir une typologie des dynamiques spatiales du marché immobilier parisien.

4.3.2 *Résultats*

Préliminaire

We show in Fig. 9 typologies of temporal transactional profiles for total stocks. Temporal dynamics show different reactions of local territories to the 2008 crisis, in particular a strong differentiation between urban and rural areas. More precise classification into urban territories are still to be investigated when the analysis will be pushed further.

4.3.3 *Une Stratégie pour étudier les signes précurseurs de rupture de potentiels*

The span of the end of this database coincides with planification phases of the Grand Paris Express that we already mentioned. We aim to seek for early warnings of potential station implantation, in correspondance with different stages of the project, in order to verify if intrinsic territorial dynamics were already present or if the announcement of a new station induced a local phase transition.

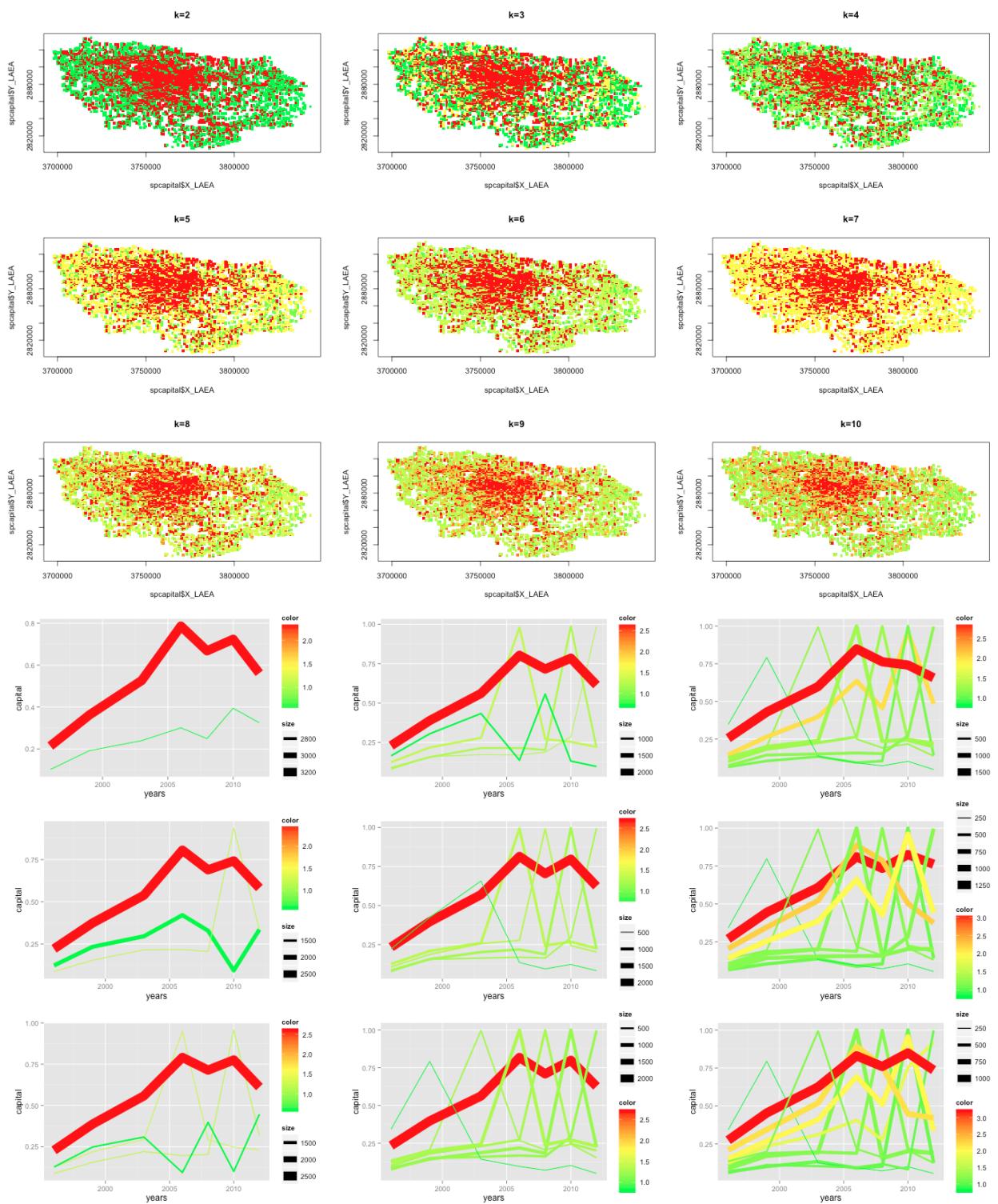


FIGURE 9 :

4.4 RELATIONS RÉSEAUX-TERRITOIRES EN AFRIQUE DU SUD

4.4.1 *Contexte*

BAFFI studied in her thesis project [11] qualitatively the role of South African railways in segregations and integration processes, aims to use an extensive database of railway growth and population dynamics in cities on the last 100 years produced during the thesis.

In particular, she showed qualitatively that dynamics between territories and networks profoundly changed at the end of the apartheid, transforming a tool of sordid planned segregation (network shaped was optimized to minimize unwanted accessibility) into an integration tool thanks to recent changes in network topology patterns.

4.4.2 *Objectifs*

We can use first the particular shape of that network to control on local and global topology effects (but this is quite equivalent as controlling on accessibility), and in a second time the historical events as statistic instruments, assuming that territorial dynamics and network dynamics responded differently to these. We expect to learn from these project informations on interactions at long time scale and large spatial scale, in a very particular context of constrained growth.

4.4.3 *Développements possibles*

The method of instruments in statistics [8] is used to identify causal relationships between variables, in a different way than Granger causality test for example. Trying to identify causalities between network dynamics and territorial dynamics is of crucial importance to test our theoretical assumption on the existence of co-evolution.

4.5 INVESTIGATION EMPIRIQUE DE L'EXISTENCE DE L'EQUILIBRE UTILISATEUR STATIQUE

L'Equilibre Utilisateur Statique est un cadre puissant pour l'étude théorique du trafic. Malgré l'hypothèse restreignant de stationnarité des flots qui intuitivement limite son application aux systèmes de trafic réels, de nombreux modèles opérationnels qui l'implémentent sont toujours utilisés sans validation empirique de l'existence de l'équilibre. Nous étudions celle-ci sur un jeu de données de trafic couvrant trois mois sur la région parisienne. L'implémentation d'une application d'exploration interactive de données spatio-temporelles permet de formuler l'hypothèse d'une forte hétérogénéité spatiale et temporelle, guidant les études quantitatives. L'hypothèse de flots localement stationnaires est invalidée en première approximation par les résultats empiriques, comme le montrent une forte variabilité spatio-temporelle des plus courts chemins et des mesures topologiques du réseau comme la centralité de chemin. De plus, le comportement de l'index d'autocorrelation spatiale pour les motifs de congestion à différentes portées spatiales suggère une évolution chaotique à l'échelle locale, en particulier lors des heures de pointe. Nous discutons finalement les implications de ces résultats empiriques et proposons des possibles développements futurs basés sur l'estimation de la stabilité dynamique au sens de Lyapounov des flots de trafic.

4.5.1 *Introduction*

La modélisation du trafic a été largement étudiée depuis les travaux séminaux de Wardrop ([257]) : les enjeux économiques et techniques justifient entre autre le besoin d'une compréhension fine des mécanismes régissant les flots de trafic à différentes échelles. Différentes approches aux objectifs différents coexistent aujourd'hui, parmi lesquels on trouve par exemple les modèles dynamiques de micro-simulation, généralement opposés aux techniques de basant sur l'équilibre. Tandis que la validité des modèles microscopiques a été étudiée de façon conséquente et leur application souvent questionnée, la littérature est relativement pauvre en études empiriques assurant l'hypothèse d'équilibre stationnaire du cadre de l'Equilibre Utilisateur Statique (EUS). De nombreux développements plus réalistes ont été documentés dans la littérature, tels l'Equilibre Utilisateur Dynamique Stochastique (EUDS) (voir pour une description par exemple [125]). A un niveau intermédiaire entre les cadres statiques et stochastiques se trouve l'Equilibre Utilisateur Stochastique Restreint, pour lequel les choix d'itinéraire des utilisateurs sont contraints à un ensemble d'alternatives réalisables ([222]). D'autres extensions prenant en compte le comportement de

l'utilisateur via des modèles de choix ont été proposés plus récemment, comme [275] qui inclut à la fois l'influence de la tarification routière et de la congestion sur le choix avec un modèle Probit. La relaxation d'autres hypothèses restrictives comme la maximisation pure de l'utilité par l'utilisateur ont aussi été introduites, tels l'Equilibre Utilisateur Borné décrit par [171]. Dans ce cadre, l'utilisateur est satisfait si son utilité tombe dans un intervalle et l'équilibre est achevé lorsque chaque utilisateur est satisfait. Les dynamiques résultantes sont plus complexes comme révélé par l'existence d'équilibres multiples, et permettent de rendre compte de faits stylisés spécifiques comme des évolutions irréversibles du réseau comme développé par [119]. D'autres modèles d'attribution de trafic inspirés d'autres domaines ont également été plus récemment proposés : dans [211], une définition étendue de la centralité de chemin qui combine linéairement la centralité des flots non-constraints avec une centralité pondérée par le temps de parcours permet d'obtenir une forte corrélation avec les flots de trafic effectifs, fournissant ainsi un modèle d'attribution de trafic. Cela fournit également des applications pratiques comme l'optimisation de la distribution spatiale des capteurs de trafic. Malgré ces nombreux développements, de nombreuses études et applications concrètes se reposent toujours sur l'Equilibre Utilisateur Statique. La région parisienne utilise par exemple un modèle statique (MODUS) pour gérer et planifier le trafic. [155] introduit un modèle statique de flots qui inclut les recherches locales et le choix du parking : il est légitime de s'interroger, en particulier à de si faibles échelles, si la stationnarité de la distribution des flots est une réalité. Une example d'exploration empirique des hypothèses classiques est donné par [277], pour lequel les choix d'itinéraires révélés sont étudiés. Les conclusions questionnent le "premier principe de Wardrop" qui implique que les utilisateurs choisissent parmi un ensemble d'alternatives parfaitement connu. Dans le même esprit, nous étudions l'existence possible de l'équilibre en pratique. Plus précisément, l'EUS suppose une distribution stationnaire des flots sur l'ensemble du réseau. Cette hypothèse reste valable dans le cas d'une stationnarité locale, tant que l'échelle temporelle d'évolution des paramètres est considérablement plus grande que les échelles typiques de voyage. Le second cas qui est plus plausible et de plus compatible avec les cadres théoriques dynamiques est testé ici.

La suite de ce travail s'organise ainsi : la procédure de collection de données ainsi que le jeu de données sont décrits ; nous présentons ensuite une application interactive pour l'exploration du jeu de données, dans le but de fournir une intuition sur les motifs présents ; puis nous donnons divers résultats d'analyses quantitatives allant dans le sens d'indices convergents pour une

non-stationnarité des flots de trafic; nous discutons finalement les implications de ces résultats et des développements possibles.

4.5.2 Collecte des données

Construction du jeu de données

Nous proposons de travailler sur l'étude de cas de la région métropolitaine de Paris. Un jeu de données ouvert a été construit, comprenant les liens autoroutiers dans la région, par collecte des données publiques en temps réel des temps de parcours (disponible sur www.sytadin.fr). Comme rappelé par [45], la disponibilité de jeux de données ouverts pour les transports est loin d'être la règle, et nous contribuons ainsi à une ouverture par la construction de notre jeu de données. La procédure de collecte de données consiste en les points suivants, exécutés toutes les deux minutes par un script python :

- récupération de la page web brute donnant les informations de trafic
- parsing du code html afin de récupérer les identifiants des liens de trafic et les temps de parcours correspondants
- insertion des liens dans une base sqlite avec le temps courant.

Le script automatisé de collection des données continue d'enrichir la base au fur et à mesure du temps, permettant des développements futurs de ce travail sur un jeu de données plus large, et une réutilisation potentielle pour des travaux scientifiques ou opérationnels. La dernière version du jeu de données au format sqlite est disponible en ligne sous une Licence Creative Commons³.

Description des données

Une granularité de deux minutes a été obtenue pour une période de trois mois (de février 2016 à avril 2016 inclus). La granularité spatiale est en moyenne de 10km, les temps de trajet étant fournis pour les liens majeurs. Le jeu de données contient 101 liens. La variable brute utilisée est le temps de trajet effectif, à partir duquel il est possible de construire la vitesse de trajet et la vitesse relative de trajet, définie comme le rapport entre temps de trajet optimal (temps de trajet sans congestion, pris comme le temps minimal sur l'ensemble des pas de temps) et le temps de trajet effectif. La congestion est construite par inversion d'un fonction BPR simple avec exposant 1, i.e. en prenant $c_i = 1 - \frac{t_{i,\min}}{t_i}$ avec t_i temps de trajet effectif dans le lien i et $t_{i,\min}$ temps de trajet minimal.

³ à l'adresse http://37.187.242.99/files/public/sytadin_latest.sqlite3

4.5.3 Méthodes	and	Résultats
----------------	-----	-----------

Visualisation des motifs spatio-temporels de congestion

Notre approche étant entièrement empirique, une bonne connaissance des motifs existants pour les variables de traffic, en particulier de leur variations spatio-temporelles, est crucial pour guider toute analyse quantitative. En s'inspirant de la littérature étudiant la validation empirique de modèles, plus précisément les techniques de *Modélisation orientée-motifs* introduites par [117], nous nous intéressons au motifs macroscopiques à des échelles temporelles et spatiales données : d'une manière équivalente aux faits stylisés qui sont dans cette approches extraits d'un système avant de tenter de le modéliser, nous devons explorer les données de manière interactive dans le temps et l'espace afin d'identifier des motifs pertinents et les échelles associées. Une application web interactive a ainsi été implémentée pour explorer les données, à l'aide des packages R shiny et leaflet⁴. Cela permet une visualisation dynamique des motifs de congestion sur l'ensemble du réseau ou dans une zone particulière grâce au zoom. L'application est accessible en ligne à l'adresse <http://shiny.parisgeo.cnrs.fr/transportation>. La Figure 10 présente une capture d'écran de l'interface. La conclusion majeure de l'exploration interactive des données est qu'une grande hétérogénéité spatiale et temporelle est la règle. Le motif temporel le plus récurrent, la périodicité journalière des heures de pointe, est perturbée pour une proportion non négligeable de jours. En première approximation, les heures creuses peuvent être approchées par une distribution localement stationnaire des flots, tandis que les heures de pointe sont trop courtes pour pouvoir impliquer la validation de l'hypothèse d'équilibre. Concernant l'espace, aucun motif spatial particulier n'émerge clairement. Cela signifie que dans le cas d'une validité de l'équilibre utilisateur statique, les méta-paramètres régissant son établissement doivent varier à des échelles temporelles plus courtes qu'un jour. Nous postulons au contraire que le système de traffic est loin de l'équilibre, en particulier pendant les heures de pointe pendant lesquelles des transitions de phase critiques à l'origine des embouteillages émergent.

Variabilité	Spatio-temporelle	des	Trajets
-------------	-------------------	-----	---------

A la suite de l'exploration interactive des données, nous proposons de quantifier la variabilité spatiale des motifs de congestion pour

⁴ le code source de l'application et des analyses est disponible sur le dépôt ouvert du projet à <https://github.com/JusteRaimbault/TransportationEquilibrium>

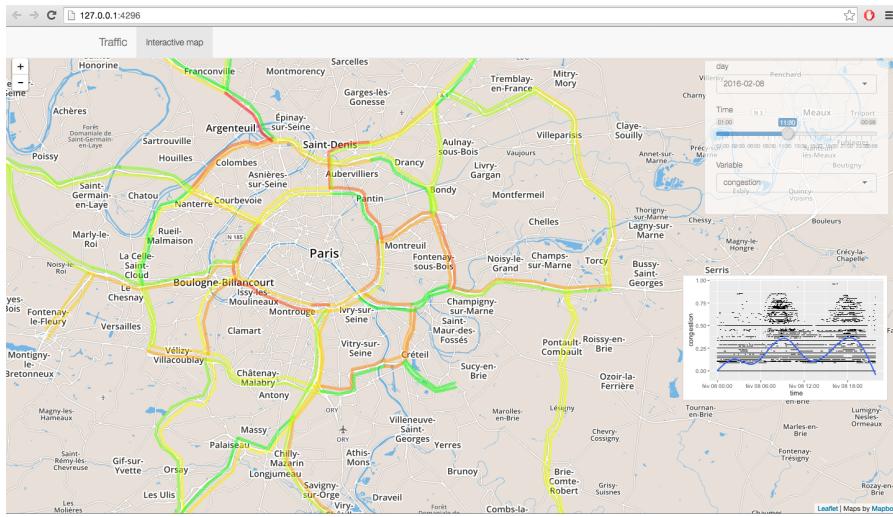


FIGURE 10 : Capture de l'application web permettant l'exploration spatio-temporelle des données de trafic pour la région Parisienne. Il est possible de choisir date et heure (précision de 15min sur un mois, réduite par rapport au jeu de données initial pour des raisons de performance). Un graphe résume les motifs de congestion pour la journée courante.

valider ou invalider l'intuition que si l'équilibre existe par rapport au temps, il est fortement dépendant de l'espace et localisé. La variabilité spatio-temporelle des plus courts chemins de trajet est une première façon d'étudier la stationnarité des flots d'un point de vue de théorie des jeux. En effet, l'Equilibre Utilisateur Statique est la distribution stationnaire des flots sous laquelle aucun utilisateur ne peut augmenter son temps de trajet en changeant son itinéraire. Une forte variabilité spatiale des plus courts chemins sur de courtes échelles spatiales révèle ainsi une non-stationnarité, puisque un même utilisateur prendra un chemin complètement différent après un court laps de temps et ne contribuera plus au même flot que précédemment. Une telle variabilité est en effet observée sur un nombre non-négligeable de chemins pour chaque jour du jeu de données. La figure 11 montre un exemple de variation spatiale extrême d'un trajet pour une paire Origine-Destination particulière. L'exploration systématique de la variabilité du temps de trajet sur l'ensemble du jeu de données, et des distances de trajet associées, confirme, comme présenté en figure , que la variation absolue du temps de trajet présente fréquemment une forte variation de son maximum sur l'ensemble des paires O-D, jusqu'à 25 minutes avec une moyenne temporelle locale autour de 10 minutes. La variabilité spatiale correspondante entraîne des détours allant jusqu'à 35km.

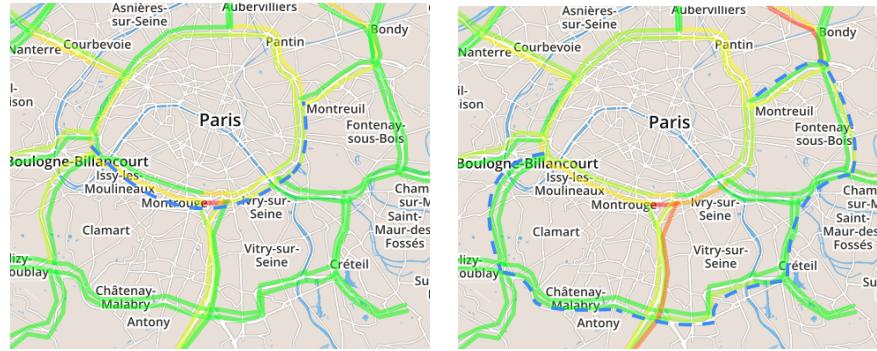


FIGURE 11 : Variabilité spatiale d'un plus court chemin en temps de trajet (trajet du plus court chemin en pointillé bleu). Dans un intervalle de seulement 10 minutes, entre le 11/02/2016 00 :06 (à gauche) et le 11/02/2016 00 :16 (à droite), le plus court chemin entre Porte d'Auteuil à l'ouest et Porte de Bagnolet à l'est, augmente en distance effective de $\simeq 37\text{km}$ (avec une augmentation du temps de trajet de seulement 6 minutes), à cause d'une forte perturbation sur le périphérique parisien.

Stabilité des mesures de réseau

La variabilité des trajectoires potentielles observée dans la section précédente peu être confirmée par l'étude de la variabilité des propriétés du réseau. En particulier, les mesures topologiques de réseau capturent les motifs globaux dans un réseau de transport. Les mesures de centralité et de connectivité des noeuds sont des indicateurs classiques pour la description des réseaux de transport comme rappelé par [25]. La littérature en transports a développé des mesures de réseau élaborées et opérationnelles, comme des mesures de robustesse pour identifier les liens critiques et mesurer la résilience globale du réseau aux perturbations (un exemple parmi d'autres est l'indice de *Robustesse du Réseau Effective* introduit dans [241]).

Plus précisément, nous étudions la centralité de chemin du réseau de transport, défini pour un noeud comme le nombre de plus courts chemins passant par celui-ci, i.e. par l'équation

$$b_i = \frac{1}{N(N-1)} \cdot \sum_{o \neq d \in V} \mathbb{1}_{i \in p(o \rightarrow d)} \quad (7)$$

où V est l'ensemble des sommets du réseau de taille N , et $p(o \rightarrow d)$ est l'ensemble des noeuds sur le plus court chemin entre les sommets o et d (le plus court chemin étant calculé avec le temps de trajet effectif). Cette mesure de centralité est plus adaptée que

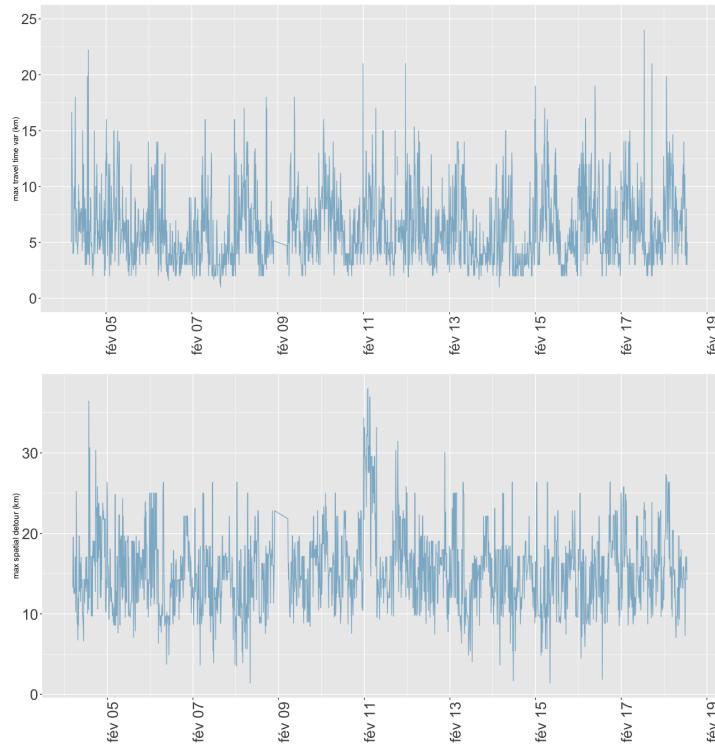


FIGURE 12 : Variabilité maximale du temps de trajet (en haut) en minutes et de la distance de trajet correspondante (en bas) pour un échantillon de deux semaines. Le graphe représente le maximum sur l'ensemble des paires Origine-Destination de la variabilité absolue entre deux pas de temps consécutifs. Les heures de pointe induisent une forte variabilité du temps de trajet, allant jusqu'à 25 minutes et une variabilité de distance jusqu'à 35km.

d'autre dans notre cas, comme la centralité de proximité qui n'inclut pas la congestion potentielle comme la centralité de chemin.

Nous montrons en Figure 4 la variation relative absolue du maximum de la centralité de chemin, pour la même fenêtre temporelle que les indicateurs empiriques précédents. Plus précisément, elle est définie par

$$\Delta b_i(t) = \frac{|\max_i(b_i(t + \Delta t)) - \max_i(b_i(t))|}{\max_i(b_i(t))} \quad (8)$$

où Δt est le pas de temps du jeu de données (la plus petite fenêtre temporelle sur laquelle une variabilité peut être捕urée). Cette variation relative absolue a une signification directe : une variation de 20% (qui est atteinte un nombre significatif de fois comme montré en Figure 13) implique dans le cas d'une variation négative, qu'au moins cette proportion de trajectoires potentielles ont changé et que la potentielle congestion locale a décrue de la même proportion. Dans le cas d'une variation positive, un seul noeud a

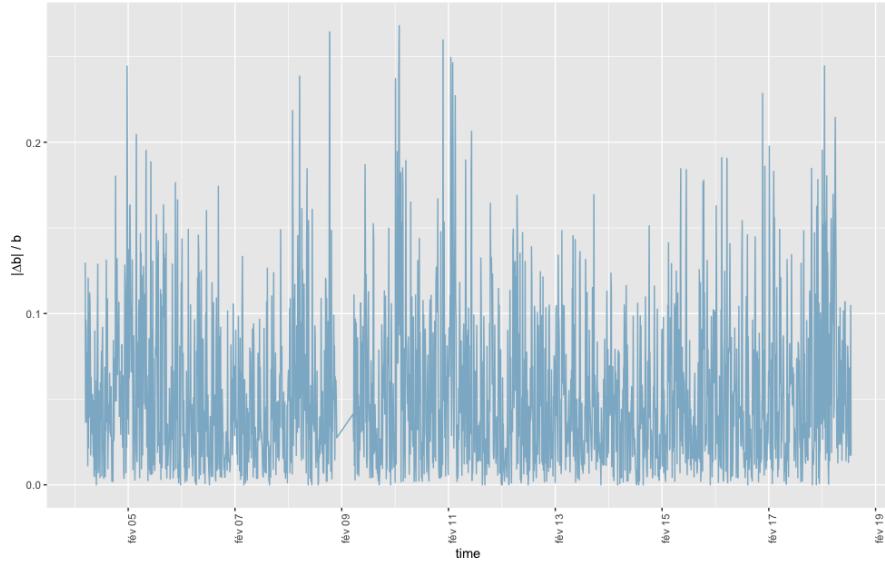


FIGURE 13 : Stabilité temporelle du maximum de la centralité de chemin. Le graphe montre dans le temps la dérivée normalisée du maximum de la centralité de chemin, qui capture ses variations relatives à chaque pas de temps. La valeur maximale de 25% correspond à de très fortes perturbations du réseau sur les liens correspondants, puisque cela implique qu'au moins cette proportion d'utilisateurs prenant le lien dans des conditions précédentes doivent prendre un trajet complètement différent.

capturé au moins 20% des trajets. Sous l'hypothèse (qu'on ne tente pas de vérifier ici et qu'on peut également supposer non vérifiée comme montré par [277], mais que l'on utilise comme un outil pour donner une intuition sur la signification concrète de la variabilité de la centralité) que les utilisateurs choisissent rationnellement le plus court chemin, et supposant que la majorité des trajets est réalisées, une telle variation de la centralité implique une variation similaire dans les flots effectifs, conduisant à la conclusion qu'ils ne peuvent être stationnaires ni dans le temps (au moins sur une échelle plus grande que Δt) ni dans l'espace.

Hétérogénéité spatiale de l'équilibre

Afin d'obtenir un point de vue différent sur la variabilité spatiale des motifs de congestion, nous proposons d'utiliser un indice d'auto-corrélation spatiale, l'indice de Moran (défini par exemple dans [246]). Utilisé plus généralement en analyse spatiale, avec diverses applications allant de l'étude de la forme urbaine à la quantification de la ségrégation, il peut être appliqué à toute variable spatiale. Il permet d'établir des relations de voisinage et révèle la consistance spatiale locale d'un équilibre s'il est appliqué à

une variable de traffic localisée. A un point donnée de l'espace, l'auto-corrélation locale pour la variable c est calculée par

$$\rho_i = \frac{1}{K} \cdot \sum_{i \neq j} w_{ij} \cdot (c_i - \bar{c})(c_j - \bar{c}) \quad (9)$$

où K est une constante de normalisation égale à la somme des poids spatiaux fois la variance de la variable et \bar{c} est la moyenne de la variable. Dans notre cas, nous choisissons des poids spatiaux de la forme $w_{ij} = \exp\left(\frac{-d_{ij}}{d_0}\right)$ avec d_0 distance typique de décroissance. L'auto-corrélation est calculée sur la congestion des liens, localisée au centre du lien. Elle capture ainsi les corrélations spatiales dans un rayon du même ordre que la distance de décroissance autour du point i . La moyenne sur l'ensemble des points fournit l'indice d'auto-corrélation spatiale I . Une stationnarité des flots devrait impliquer une stabilité temporelle de l'index.

La figure 14 présente l'évolution temporelle de l'auto-corrélation spatiale pour la congestion. Comme attendu, on observe une forte décroissance de l'auto-corrélation avec la distance de décroissance, à la fois sur l'amplitude et les moyennes temporelles. La forte variabilité temporelle implique de courtes échelles temporelles pour des fenêtres potentielles de stationnarité. Pour une distance de décroissance de 1km, en comparant l'auto-corrélation à la congestion (ajustée à l'échelle du graphe pour lisibilité), on observe que les fortes corrélations coïncident avec les heures creuses, tandis que les heures de pointe correspondent à une décroissance des corrélations.

Notre interprétation, combinée avec la variabilité observée des motifs spatiaux, est que les heures de pointe correspondent à un comportement chaotique du système, puisque les bouchons peuvent émerger dans n'importe quel lien du réseau : la corrélation disparaît alors puisque l'espace des phases atteignables pour un système dynamique chaotique est rempli uniformément par les trajectoires, de façon équivalente à des vitesses relatives qui apparaîtraient comme aléatoires et indépendantes.

4.5.4 Discussion

Implications théoriques et pratiques des conclusions empiriques

Nous prétendons que les implications théoriques de ces résultats empiriques n'impliquent pas nécessairement un rejet total du cadre de l'Equilibre Utilisateur Statique, mais révèlent plutôt un besoin de plus fortes connexions entre la littérature théorique et les études empiriques. Si chaque nouveau cadre théorique introduit est généralement testé sur un cas ou plus, il n'existe pas de comparaisons systématiques de chacun sur des jeux de données de

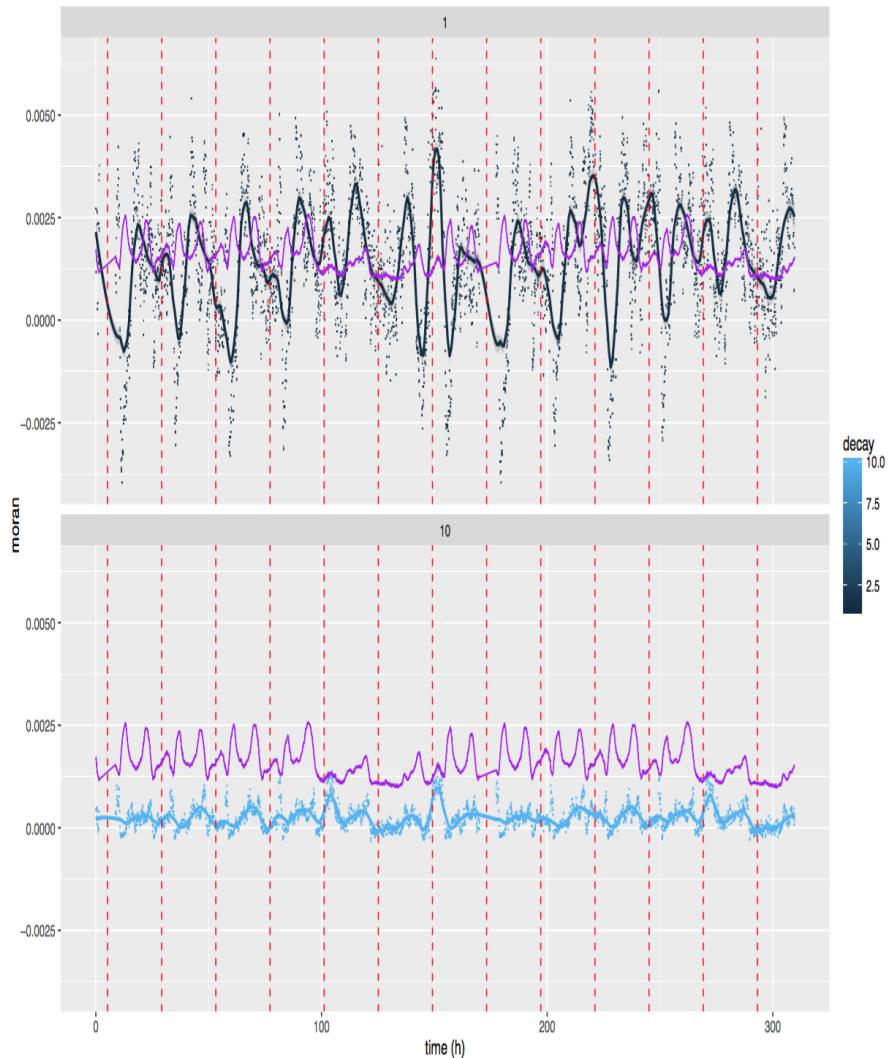


FIGURE 14 : Auto-corrélations spatiales pour les vitesses relatives sur deux semaines. Le graphe montre les valeurs de l'auto-corrélation dans le temps, pour des valeurs variables (1,10km) de la distance de décroissance. les valeurs intermédiaires de la distance de décroissance donnent une déformation relativement continue entre ces deux extrêmes. Les points sont lissés sur une fenêtre temporelle de 2h pour faciliter la lecture. Les lignes pointillées verticales correspondent à minuit de chaque jour. La courbe violette donne la vitesse relative, ajustée à l'échelle pour établir la correspondance entre les heures de pointe et les variations de l'auto-corrélation.

grande taille et variés, et pour des objectifs d'application différents (prédiction du traffic, reproduction de faits stylisés, etc.), à l'image des revues systématiques qui sont la règle en évaluation thérapeutique par exemple. Cela implique cependant des pratiques de partage des données et des modèles plus larges que celles existant couramment. La connaissance précise des potentialités d'application d'un cadre donné peut induire des développements inattendus comme l'intégration dans des modèles plus larges. L'exemple des études des interaction entre Transport et Usage du Sol (modèles *LUTI*) est une bonne illustration d'un cas où le EUS peut toujours être utilisé avec des motivations plus larges que la modélisation du traffic. [145] décrit deux modèles *LUTI*, dont l'un inclut deux équilibres pour les modèles de transport à quatre temps et pour l'évolution de l'usage du sol (localisation des ménages et emplois), l'autre étant dynamique. La conclusion est que chaque modèle a ses avantages au regard de l'objectif poursuivi, et que le modèle statique peut être utilisé pour comparer des politiques sur le temps long, tandis que le modèle dynamique fournit de l'information plus précise à de plus petites échelles temporelles. Dans le premier cas, un module de transport plus compliqué aurait été plus difficile à inclure, ce qui est un avantage du EUS dans ce cas.

Concernant les applications pratiques, il semble naturel que les modèles statiques ne devraient pas être utilisés pour la prédiction et la gestion du traffic sur de petites échelles temporelles (semaine ou jour) et que des efforts doivent être faits pour implémenter des modèles plus réalistes. Cependant, l'utilisation des modèles par la communautés des ingénieurs et des planificateurs n'est pas directement reliée aux enjeux académiques et à l'état de l'art dans le domaine. Dans le cas particulier de la France et des modèles de mobilité, [67] a montré que les ingénieurs allaient jusqu'au point de construire des problèmes inexistant et d'implémenter les modèles correspondants qu'ils avaient importé d'un contexte géographique totalement différent (la planification aux Etats-Unis). L'utilisation d'un cadre ou d'un type de modèle a des raisons historiques qui peuvent être difficiles à surmonter.

Vers des interprétations de la non-stationnarité

Une hypothèse qu'on peut formuler concernant l'origine de la non-stationnarité des flots dans le réseau, au regard de l'exploration des données et des analyses quantitatives, est que le réseau est au moins la moitié du temps fortement congestionné et dans un état critique. Les heures creuses sont les plus grandes fenêtres temporelles potentielles de stationnarité spatiale et temporelle, mais couvre moins de la moitié du temps. Comme déjà interprété dans le comportement de l'indicateur d'auto-corrélation, un comportement chaotique pourrait être à l'origine d'une telle variabilité lors des

heures congestionnées. A la manière d'un fluide supercritique qui condense sous une perturbation externe infinitésimale, l'état d'un lien peut qualitativement changer par un petit incident, produisant une perturbation du réseau qui se propage et peut même s'amplifier.

L'effet direct des événements du traffic (incidents signalés ou accidents) ne peut pas être étudié sans source de données extérieure, et un enrichissement de la base de données dans cette direction pourrait être intéressante. Cela permettrait d'établir la proportion de perturbations qui paraissent avoir un effet direct et quantifier un niveau de caractère critique de la congestion du réseau dans le temps, ou d'étudier plus précisément des phénomènes localisés comme les conséquences d'un incident de traffic sur la voie opposée.

Développements

Le travail futur pourra être planifié dans la direction d'une étude raffinée de la stabilité temporelle sur des zones du réseau, i.e. l'étude quantitative précise de la non-stationnarité des heures de pointes découverte ci-dessus. Pour cela nous proposons de calculer numériquement la stabilité de Liapounov du système dynamique régissant les flots de traffic, par l'intermédiaire d'algorithmes numériques comme ceux décrits par [113]. La valeur des exposants de Liapounov fournit l'échelle de temps sur laquelle le système instable s'éloigne de l'équilibre. Leur comparaison avec la durée des heures de pointe et le temps de trajet moyen, sur différentes zones spatiales et différentes échelles, devrait fournir plus d'information sur une possible validité de l'hypothèse de stationnarité locale. Cette technique a déjà été introduite à une autre échelle dans les études de transport, comme e.g. [245] qui étudie la stabilité des modèles de régulation de vitesse à l'échelle microscopique pour éviter l'émergence de congestion.

D'autres directions de recherche peuvent consister en le test des autres hypothèses du EUS (comme le choix rationnel du plus court chemin, qui serait cependant difficile à tester à un tel niveau d'agrégation, impliquant l'utilisation de modèles de simulation calibrés et cross-validés sur le jeu de données pour comparer différentes hypothèses, sans toutefois nécessairement une validation ou invalidation directe de l'hypothèse), ou le calcul empirique des paramètres dans les cadres d'Equilibre Utilisateur Stochastique ou Dynamique.

Conclusion

Nous avons décrit une étude empirique ayant pour but une étude simple, mais selon notre point de vue nécessaire, de l'existence de l'équilibre utilisateur statique, plus précisément de sa stationnarité

dans le temps et l'espace pour un réseau routier métropolitain principal. Un jeu de données de congestion du trafic est construite par collection de données, pour le réseau du Grand Paris sur 3 mois avec une granularité temporelle de 2 minutes. L'exploration interactive du jeu de données via une application web permettant la visualisation spatio-temporelle aide à guider les analyses quantitatives. La variabilité spatio-temporelle des plus courts chemins et de la topologie du réseau, en particulier la centralité de chemin, révèle que l'hypothèse de stationnarité ne tient généralement pas, ce qui est confirmé par l'étude de l'auto-corrélation spatiale de la congestion du réseau. Nous suggérons que nos résultats soulignent un besoin général de plus grandes connexions entre les études théoriques et empiriques, puisque cette étude permet de chasser les incompréhensions théoriques sur l'Equilibre Utilisateur Statique, et guider le choix d'application potentielles.

5

MODELING

Do or do not. There is no try.

- YODA

One does not simply *try* to model something. On that point personal experience confirms indeed that point, as I remember as an early Master student giving in to the call of incautious agent-based modeling, naively thinking that integrated models of any aspect of an urban system could be constructed, producing numerous NetLogo code lines to build a gaz factory with unfounded internal processes, an extremely poor external validation and no internal validation. This was a try and therefore a step towards the dark side of models bricolage. The construction of a computational model of simulation is a rigorous exercise that one can not improvise, as much as statistical modeling. Recent progresses in the field [12] help to that purpose, and modular model construction and validation is one tool useful to avoid becoming lost in shady places.

We propose in this chapter simple modeling experiments, conceived to be preliminaries for more elaborated tests of our theory. We begin with a simple diffusion-aggregation model of urban growth as a relatively small scale. Beginning with simple assumptions does not mean a non-rigorous exploration of the model, that is therefore explored and calibrated on real data. The fact that we reproduce existing urban forms without the use of networks suggest either the total absence of network influence at this scale, or its very strong influence yielding apparent random effects that disappear in average calibration. We propose then to simply couple this model with a network generation heuristic in order to study feasible correlations between morphology and network. The absence of coupled calibration avoids to draw empirical conclusion but the method is satisfying in itself as it permits the generation of synthetic territorial configurations where correlation structure is controlled. We finally describe a project of benchmark of diverse heuristic models for network generation.

5.1 UN MODÈLE SIMPLE DE CROISSANCE URBAINE

TODO : insert / translate Density paper

5.2 GÉNÉRATION DE CONFIGURATIONS TERRITORIALES CORRÉLÉES

This section aims to explore the sequential coupling between previous model of density generation and an heuristic of network growth. We explore therein the feasible space of correlations between network measures and morphological measures.

5.2.1 Données Géographiques corrélées de Densité et de Réseau

Contexte

En géographie, l'utilisation de données synthétiques est plus généralement axée vers l'utilisation de population synthétiques au sein de modèles basés agents (mobilité, modèles *LUTI*) [201]. On peut également citer des méthodes d'analyse spatiales qui s'en rapprochent : par exemple, l'extrapolation d'un champ spatial continu à partir d'un échantillon discret, par une estimation par noyaux par exemple, peut être compris comme la génération d'un jeu de données synthétiques (même si ce n'est pas le point de vue initial, comme pour la Regression Géographique Pondérée [51], dans laquelle les noyaux de taille variables n'interpolent pas des données au sens propre mais extrapolent des variables abstraites représentant l'interaction entre variables explicites). Dans le domaine de la modélisation en géographie quantitative, dans le cas de *modèles jouets* ou de modèles hybrides, une configuration initiale cohérente est souvent essentielle : un ensemble de configurations initiales possibles est alors un jeu de données synthétiques sur lesquelles le modèle est testé : le premier modèle Simpop [230], pionnier d'une famille de modèles par la suite paramétrisés par des données réelles, pourrait rentrer dans ce cadre mais était lancé sur une spatialisation synthétique unique. De même, il a été souligné la difficulté de générer une configuration initiale pour une infrastructure de transport dans le cas du modèle SimpopNet [233], alors qu'il s'agit un point essentiel dans la connaissance du comportement du modèle. Il a récemment été proposé de contrôler systématiquement les effets de la configuration spatiale sur le comportement de modèles de simulation spatialisés [74], méthodologie pouvant être interprétée comme un contrôle par données statistiques spatiales. L'enjeu est de pouvoir alors distinguer effets propres dus à la dynamique intrinsèque du modèle, d'effet particuliers dus à la structure géographique du cas d'application. Celui-ci est crucial pour la validation des conclusions issues des pratiques de modélisation et simulation en géographie quantitative.

Formalisation

Dans notre cas, nous proposons de générer des systèmes de villes représentés par une densité spatiale de population $d(\vec{x})$ et la donnée d'un réseau de transport $n(\vec{x})$, représenté de façon simplifiée, pour lesquels on serait capable de contrôler les correlations entre mesures morphologiques de la densité urbaine et caractéristiques du réseau. La question de l'interaction entre territoire et réseaux de transport est un sujet d'étude classique [191] mais extrêmement complexe et difficile à quantifier [190]. Une modélisation dynamique des processus impliqués devrait apporter des connaissances sur ces interactions ([46], p. 162-163). Dans ce cadre, nous développons un couplage *simple* (c'est à dire sans boucle de rétroaction) entre un modèle de morphogenèse urbaine et un modèle de génération de réseau.

MODÈLE DE DENSITÉ Les modèle de densité est celui décrit et exploré dans la section précédente. Nous l'utilisons pour la génération conditionnelle du réseau.

MODÈLE DE RÉSEAU D'autre part, on est capable de générer par un modèle N un réseau de transport planaire à une échelle équivalente, étant donné une distribution de densité. La génération du réseau étant conditionnée à la donnée de la densité, les estimateurs des indicateurs de réseau seront conditionnels d'une part, et d'autre part les formes urbaines et du réseau devraient nécessairement être corrélées, les processus n'étant pas indépendants. La nature et la modularité de ces correlations selon la variation des paramètres des modèles restent à déterminer par l'exploration du modèle couplé.

La procédure de génération heuristique de réseau est la suivante :

1. Un nombre fixé N_c de centres qui seront les premiers noeuds du réseau est distribué selon la distribution de densité, suivant une loi similaire à celle d'agrégation, i.e. la probabilité d'être distribué sur une case est $\frac{(P_i/P)^\alpha}{\sum(P_i/P)^\alpha}$. La population est ensuite répartie selon les zones de Voronoi des centres, un centre cumulant la population des cases dans son emprise.
2. Les centres sont connectés de façon déterministe par percolation entre plus proches clusters : tant que le réseau n'est pas connexe, les deux composantes connexes les plus proches au sens de la distance minimale entre chacun de leurs sommets sont connectées par le lien réalisant cette distance. On obtient alors un réseau arborescent.
3. Le réseau est alors modulé par ruptures de potentiels afin de se rapprocher de formes réelles. Plus précisément, un potentiel

d'interaction gravitaire généralisé entre deux centres i et j est défini par

$$V_{ij}(d) = \left[(1 - k_h) + k_h \cdot \left(\frac{P_i P_j}{P^2} \right)^\gamma \right] \cdot \exp \left(-\frac{d}{r_g(1 + d/d_0)} \right)$$

où d peut être la distance euclidienne $d_{ij} = d(i, j)$ ou la distance par le réseau $d_N(i, j)$, $k_h \in [0, 1]$ un poids permettant de changer le rôle des population dans le potentiel, γ régissant la forme de la hiérarchie selon les valeurs des populations, r_g distance caractéristique de décroissance et d_0 paramètre de forme.

4. Un nombre $K \cdot N_L$ de nouveaux liens potentiels est pris comme les couples ayant le plus grand potentiel pour la distance euclidienne ($K = 5$ est fixé).
5. Parmi les liens potentiels, N_L sont effectivement réalisés, qui sont ceux ayant le plus faible rapport $V_{ij}(d_N)/V_{ij}(d_{ij})$: à cette étape seul l'écart entre distance euclidienne et distance par le réseau compte, ce rapport ne dépendant plus des populations et étant croissant en d_N à d_{ij} fixé.
6. Le réseau est planarisé par création de noeuds aux intersections éventuelles créées par les nouveaux liens.

Notons que la construction du modèle de génération est heuristique, et que d'autres types de modèles comme un réseau biologique auto-généré [TeroAl10], une génération par optimisation locale de contraintes géométriques [19] ou un modèle de percolation plus complexe que celui utilisé, peuvent le remplacer. Ainsi, dans le cadre d'une architecture modulaire où le choix entre différentes implémentations d'une brique fonctionnelle peut être vue comme méta-paramètre [72], on pourrait choisir la fonction de génération adaptée à un besoin donné (par exemple proximité à des données réelles, contraintes sur les relations entre indicateurs de sortie, variété de formes générées, etc.).

ESPACE DES PARAMÈTRES L'espace des paramètres du modèle couplé¹ est constitué des paramètres de génération de densité $\vec{\alpha}_D = (P_m/N_G, \alpha, \beta, n_d)$ (on s'intéresse pour simplifier au rapport entre population et taux de croissance, i.e. le nombre d'étapes nécessaires pour générer) et des paramètres de génération de réseau $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. On notera $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

¹ Le couplage faible permet de limiter le nombre total de paramètres puisqu'un couplage fort incluant des boucles de retroaction comprendrait nécessairement des paramètres supplémentaires pour régler la forme et l'intensité de celles-ci. Pour espérer le diminuer, il faudrait concevoir un modèle intégré, ce qui est différent d'un couplage fort dans le sens où il n'est pas possible de figer l'un des sous-systèmes pour obtenir un modèle de l'autre correspondant au modèle non-couplé.

INDICATEURS On quantifie la forme urbaine et la forme du réseau, dans le but de moduler la corrélation entre ces indicateurs.

La forme est définie par un vecteur $\vec{M} = (r, \bar{d}, \varepsilon, a)$ donnant auto-corrélation spatiale (indice de Moran), distance moyenne, entropie, hiérarchie (voir [153] pour une définition précise de ces indicateurs). Les mesures de la forme du réseau $\vec{G} = (\bar{c}, \bar{l}, \bar{s}, \delta)$ sont, avec le réseau noté (V, E) ,

- Centralité moyenne \bar{c} , définie comme la moyenne de la *betweenness-centrality* (normalisée dans $[0, 1]$) sur l'ensemble des liens.
- Longueur moyenne des chemins \bar{l} définie par $\frac{1}{d_m} \frac{2}{|V| \cdot (|V|-1)} \sum_{i < j} d_N(i, j)$ avec d_m distance de normalisation prise ici comme la diagonale du monde $d_m = \sqrt{2}N$.
- Vitesse moyenne [15], qui correspond à la performance du réseau par rapport au trajet à vol d'oiseau, définie par $\bar{s} = \frac{2}{|V| \cdot (|V|-1)} \sum_{i < j} \frac{d_{ij}}{d_N(i, j)}$.
- Diamètre du réseau $\delta = \max_{ij} d_N(i, j)$

COVARIANCE ET CORRELATION On s'intéressera à la matrice de covariance croisée $\text{Cov}[\vec{M}, \vec{G}]$ entre densité et réseau, estimée sur un jeu de n réalisations à paramètres fixés $(\vec{M}[D(\vec{\alpha})], \vec{G}[N(\vec{\alpha})])_{1 \leq i \leq n}$ par l'estimateur standard non-biaisé. On prend comme corrélation associée la corrélation de Pearson estimée de la même façon.

Implémentation

Le couplage des modèles génératifs est effectué à la fois au niveau formel et au niveau opérationnel, c'est à dire qu'on fait interagir des implémentations indépendantes. Pour cela, le logiciel OpenMole [223] utilisé pour l'exploration intensive, offre le cadre idéal de par son langage modulaire permettant de construire des *workflows* par composition de tâches à loisir et de les brancher sur divers plans d'expérience et sorties. Pour des raisons opérationnelles, le modèle de densité est implémenté en langage scala comme un plugin d'OpenMole, tandis que la génération de réseau est implémentée en langage basé-agent NetLogo [264], ce qui facilite l'exploration interactive et construction heuristique interactive. Le code source est disponible pour reproductibilité sur le dépôt du projet².

Résultats

L'étude du modèle de densité seul est développée dans [215]. Il est notamment calibré sur les données de la grille européenne de densité, sur des zones de 50km de côté et de résolution 500m pour

² à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic>

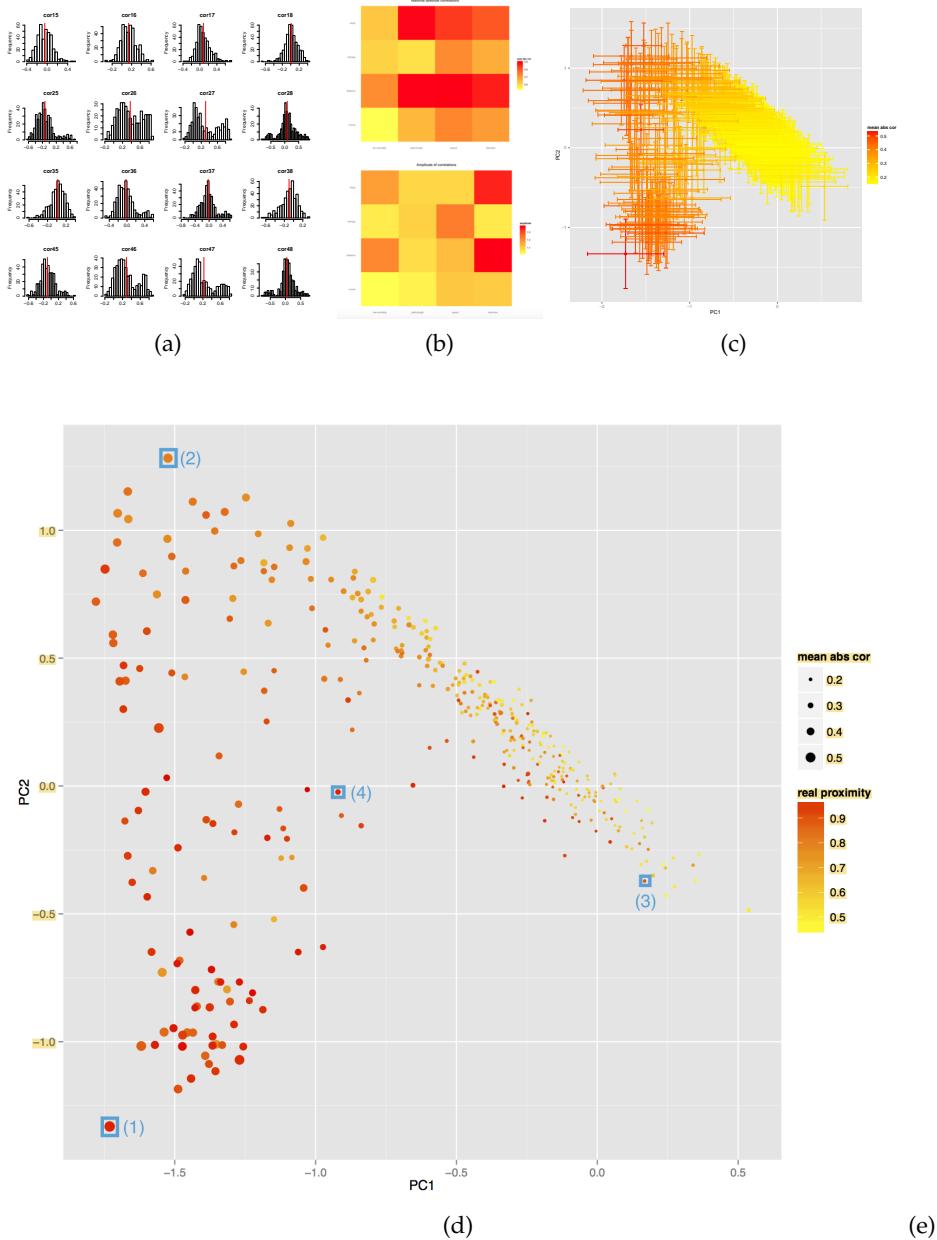


FIGURE 15 :

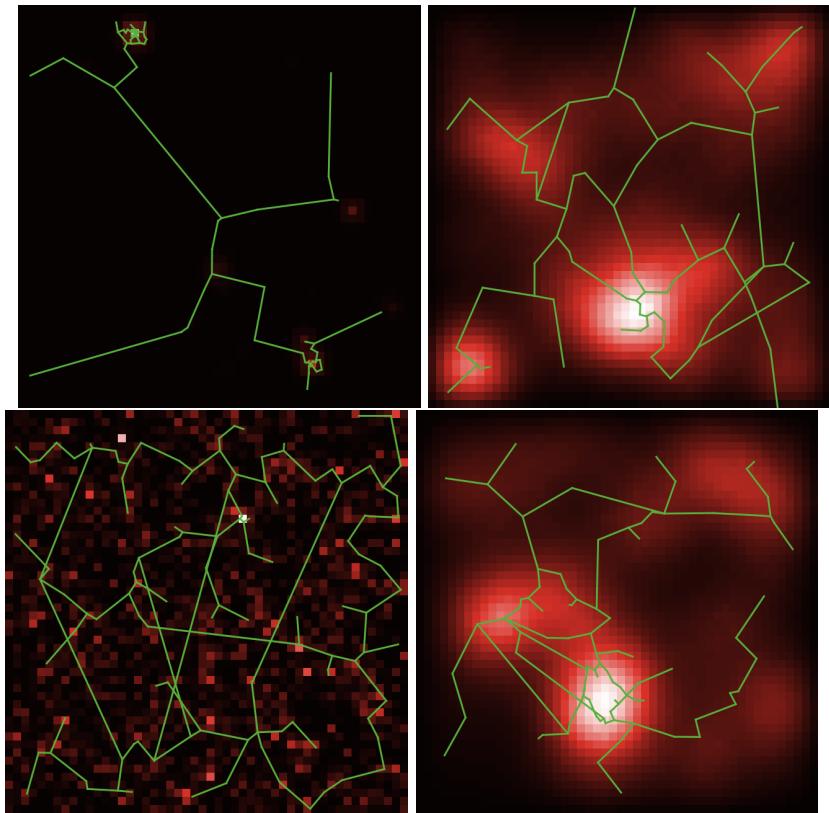


FIGURE 16 :

lesquelles les valeurs réelles des indicateurs ont été calculées pour l'ensemble de l'Europe. D'autre part, une exploration brutale du modèle permet d'estimer l'ensemble des sorties possibles dans des bornes raisonnables pour les paramètres (grossièrement $\alpha \in [0.5, 2]$, $N_G \in [500, 3000]$, $P_m \in [10^4, 10^5]$, $\beta \in [0, 0.2]$, $n_d \in \{1, \dots, 4\}$). La réduction à un plan de l'espace des objectif par une Analyse en Composantes Principales (variance expliquée à deux composantes $\simeq 80\%$) permet d'isoler un nuage de points de sorties recouvrant assez fidèlement le nuage des points réels, ce qui veut dire que le modèle est capable de reproduire morphologiquement l'ensemble des configurations existantes.

A densité donnée, l'exploration de l'espace des paramètres du modèle de réseau suggèrent une assez bonne flexibilité sur des indicateurs globaux \vec{G} , ainsi que de bonnes propriétés de convergence. Pour une étude du comportement précis, voir l'appendice donnant les regressions traduisant le comportement du modèle couplé. Dans le but d'illustrer la méthode de génération de données synthétiques, l'exploration a été orientée vers l'étude des correlations.

Etant donné la grande dimension relative de l'espace des paramètres, une exploration par grille exhaustive est impossible. On utilise un plan d'expérience par criblage (hypercube latin), avec les

bornes indiquées ci-dessus pour $\vec{\alpha}_D$ et pour $\vec{\alpha}_N$, on a $N_C \in [50, 120]$, $r_g \in [1, 100]$, $d_0 \in [0.1, 10]$, $k_h \in [0, 1]$, $\gamma \in [0.1, 4]$, $N_L \in [4, 20]$.

Concernant le nombre de réplications du modèle pour chaque valeur des paramètres, moins de 50 sont nécessaires pour obtenir sur les indicateurs des intervalles de confiance à 95% de taille inférieure aux déviations standard. Pour les correlations, une centaine donne des IC (obtenus par méthode de Fisher) de taille moyenne 0.4, on fixe donc $n = 80$ pour l'expérience. La figure 15 donne le détail des résultats de l'exploration. On retiendra les résultats marquants suivants au regard de la génération de données synthétiques corrélées :

- les distributions empiriques des coefficients de correlations entre indicateurs de forme et indicateurs de réseaux ne sont pas simples, pouvant être bimodales (par exemple $\rho_{46} = \rho[r, \bar{l}]$ entre l'index de Moran et le chemin moyen).
- On arrive à générer un assez haut niveau de correlation pour l'ensemble des indicateurs, la correlation absolue maximale variant entre 0.6 et 0.9; l'amplitude varie quant à elle entre 0.9 et 1.6, ce qui permet un large spectre de valeurs. L'espace couvert dans un plan principal a une étendue certaine mais n'est pas uniforme : on ne peut pas moduler à loisir n'importe quel coefficients, ceux-ci étant liés par les processus de génération sous-jacent. Une étude plus fine aux ordres suivants (corrélation des correlations) serait nécessaire pour cerner exactement la latitude dans la génération.
- les points les plus corrélés en moyenne sont également ceux les plus proches des données réelles, ce qui confirme l'intuition d'une forte interdépendance en réalité.
- Des exemples concrets pris sur des points particuliers distants dans le plan principal montre que des configurations de densité proches peuvent présenter des profils de correlations très différents.

Développements

Il est possible de raffiner cette étude en étendant la méthode de contrôle des correlations. La connaissance très fine du comportement de N (distribution statistiques sur une grille fine de l'espace des paramètres) conditionnée à D devrait permettre de déterminer exhaustivement $N^{<-1>}|D$ et avoir plus de latitude dans la génération des correlations. On pourra également appliquer des algorithmes spécifiques d'exploration pour essayer atteindre des configurations exceptionnelles réalisant un niveau de corrélation voulu, ou au moins pour découvrir l'espace des correlations atteignables par la méthode de génération [63].

5.2.2 Discussion

Positionnement

Scientifique

Notre démarche s'inscrit dans un cadre épistémologique particulier. En effet, d'une part la volonté de multi-disciplinarité et d'autre part l'importance de la composante empirique couplée aux méthodes d'exploration computationnelles, en font une approche typique des sciences de la complexité, comme le rappelle la structure de la feuille de route pour les systèmes complexes [43] qui croise des grandes questions transversales aux disciplines à une intégration verticale de celles-ci, qui implique la construction de modèles multi-échelles hétérogènes présentant souvent les aspects précédent. Le croisement de connaissances empiriques issues de la fouille de données avec celles issues de la simulation est souvent central dans leur conception ou leur exploration, et les résultats présentés ici en sont un exemple typique pour le cas de l'exploration.

Applications

Directes

En partant du deuxième exemple, qui s'est arrêté à la génération des données synthétiques, on peut proposer des pistes d'application directe qui donneront un aperçu de l'éventail des possibilités.

- La calibration de la composante de génération de réseau, à densité donnée, sur des données réelle de réseau de transport (typiquement routier vu les formes heuristiques obtenues, il devrait par exemple être aisément d'utiliser les données ouvertes d'OpenStreetMap³ qui sont de qualité raisonnable pour l'Europe, du moins pour la France [110], avec toutefois des ajustements à faire sur le modèle pour supprimer les effets de bord du à sa structure, par exemple en le faisant générer sur une surface étendue pour ne garder qu'une zone centrale sur laquelle la calibration aurait lieu) permettrait en théorie d'isoler un jeu de paramètres représentant fidèlement des situations existantes à la fois pour la forme urbaine et la forme du réseau. Il serait alors possible de dériver une "corrélation théorique" pour celles-ci, étant donné qu'une corrélation empirique n'est en théorie pas calculable puisqu'une seule instance des processus stochastiques est observée. Vu la non-ergodicité des systèmes urbains [207], il y a de fortes chances pour que ces processus soient différents d'une zone géographique à l'autre (ou selon un autre point de vue qu'ils soient dans un autre état des meta-paramètres, dans un autre régime) et que leur interprétation en tant que réalisations d'un même processus stochastique n'ait aucun sens, entraînant l'impossibilité du calcul des covariations. En attribuant un

³ <https://www.openstreetmap.org>

jeu de données synthétiques similaire à une situation donnée, on serait capable de calculer une sorte de *correlation intrinsèque* propre à la situation, qui émerge en fait en réalité des interdépendances temporelles des composantes. Connaitre celle-ci renseigne alors sur ces interdépendances, et donc sur les relations entre réseaux et territoires.

- Comme déjà évoqué, la plupart des modèles de simulation nécessitent un état initial, généré artificiellement à partir du moment où la paramétrisation n'est pas effectuée totalement à partir de données réelles. Une analyse de sensibilité avancée du modèle implique alors un contrôle sur les paramètres de génération du jeu de données synthétique, vu comme méta-paramètre du modèle [74]. Dans le cas d'une analyse statistique des sorties du modèle, on est alors capable d'effectuer un contrôle statistique au second ordre.
- On a étudié des processus stochastiques dans le premier exemple, au sens de séries temporelles aléatoires, alors que le temps ne jouait pas de rôle dans le second. On peut suggérer un couplage fort entre les deux composantes du modèle (ou la construction d'un modèle intégré) et observer les indicateurs et correlations à différents pas de temps de la génération. Dans le cas d'une dynamique, de par les rétroactions, on a nécessairement des effets de propagation et donc l'existence d'interdépendances décalées dans l'espace et le temps [196], étendant le domaine d'étude vers une meilleure compréhension des corrélations dynamiques.

Généralisation

On s'est limité au contrôle des premiers et second moments des données générées, mais il est possible d'imaginer une généralisation théorique permettant le contrôle des moments à un ordre arbitraire. Toutefois, la difficulté de génération dans un cas concret complexe, comme le montre l'exemple géographique, questionne la possibilité de contrôle aux ordres supérieurs tout en gardant un modèle à la structure cohérente au nombre de paramètres relativement faibles. Par contre, l'étude de structures de dépendances non-linéaires comme celles utilisées dans [64] est une piste de développement intéressante.

5.2.3 Conclusion

On a ainsi proposé une méthode abstraite de génération de données synthétiques corrélées à un niveau contrôlé. Son implantation partielle dans deux domaines très différents montre sa flexibilité et l'éventail des applications potentielles. De manière générale, il est

essentiel de généraliser de telles pratiques de validation systématique de modèles par étude statistique, en particulier pour les modèles agents pour lesquels la question de la validation reste encore relativement ouverte.

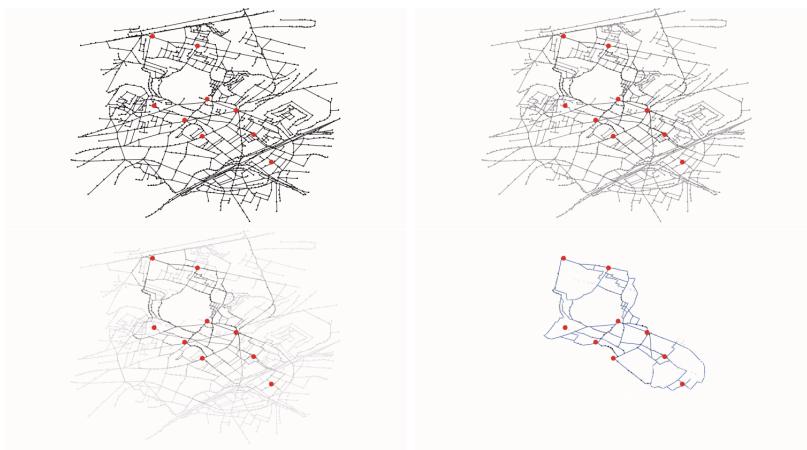


FIGURE 17 :

5.3 MODÈLES DE CROISSANCE DE RÉSEAU

5.3.1 Comparer les heuristiques de croissance de réseau

Pour la croissance du réseau en tant que tel, de nombreuses heuristiques existent pour générer un réseaux sous certaines contraintes. Comme déjà développé précédemment, des modèles économiques de croissance de réseau au heuristiques d'optimisation locale, aux mécanismes géographiques ou à la croissance de réseau biologique, chacun a ses avantages et particularités propres. Un travail futur aura pour but de comparer ces diverses méthodes contre les valeurs réelles des indicateurs pour le réseau de routes européen. La Fig. 17 présente un travail préliminaire présenté dans [218] qui explore des applications des modèles de croissance de réseau biologique. D'autre part, comme présenté dans la section sur la reproductibilité, des modèles d'optimisation locale ont également été testés.

5.3.2 Vers des modèles simples de morphogenèse de réseau

An interdisciplinary project that was just launched with a Physicist LAGESSE, an Architect HACHI and a Computer Scientist DUGUE aims at finding consistent models of urban street network morphogenesis, regarding urban design particularities, geographical rules and complex network indicators feedbacks. Models of network morphogenesis were already discuss here and the aim of this project is to gain insight from the interdisciplinary vision to explore the potentiality of such models. In the frame of our thesis, it is logically situated within the morphogenesis theoretical part and network growth modeling heuristics.

6

TRANSPORTATION GOVERNANCE MODELING

This single section chapter is differentiated from the previous one as it makes a step further towards more complex models. A toy-model introducing governance processes is described. Such exploration logically enters our theoretical framework to try to validate or invalidate the network necessity assumption : if non-linear necessary processes are highlighted and validated against stylized facts, it argues towards the validation of this assumption.

Other targeted projects such as the exploration of an hybrid macro-economic/accessibility-based model to explore transportation companies line implementation strategies are still at the state of ideas and are not described here.

6.1 LE

MODÈLE

LUTECIA

TODO : *insert and translate Lutecia paper*

Troisième partie

SYNTHESIS

This concluding remark, for now a brief roadmap, is one objective of our thesis as implementation of our theory and thus is expected to become a consequent part. We conclude here this preliminary work by perspectives and roadmap. This part make the synthesis of what was build until now, towards a delicate though robust edifice.

A ROADMAP FOR AN OPERATIONAL FAMILY OF MODELS OF COEVOLUTION

As previously stated, one of our principal aims is the validation of the network necessity assumption, that is the differentiating point with a classic evolutive urban theory. To do so, toy-model exploration and empirical analysis will not be enough as hybrid models are generally necessary to draw effective and well validated conclusions. We briefly give an overview of planned work in the following, that will be the conclusion of this Memoire.

7.1 OBJECTIVES

We expect to product *models of coevolution*, with the emphasis on processes of coevolution, to directly confront the theory. They will be necessary a flexible family because of the variety of scales and concrete cases we can include and we already began to explore in preliminary studies. Processes already studied can serve either as a thematic bases for a reuse as building bricks in a multi-modeling context, or as methodological tools such as synthetic data generator for synthetic control. Finally, we mean by operational models hybrid models, in the sense of semi-parametrized or semi-calibrated on real datasets or on precise stylized facts extracted from these same datasets. This point is a requirement to obtain a thematic feedback on geographical processes and on theory.

7.2 CAS

D'ÉTUDE

Currently we expect to work on the following case studies to build these hybrid models :

- Dynamical data for Bassin Parisien should allow to parametrize and calibrate a model at this temporal and spatial scale.
- On larger scales, South African dataset of BAFFI will along empirical analysis also be used to parametrize hybrid co-evolution models.
- A possibility that is not currently set up (and that may however be difficult because of a disturbing closed-data policy among a frightening large number of scientists!) is the exploitation of French railway growth dataset (with population dataset) used in [46], that would also provide an interesting case study on other regimes, scales and transportation mode.

7.3 FEUILLE DE ROUTE

We give the following (non-exhaustive and provisory) roadmap for modeling explorations (theoretical and empirical domains being still explored conjointly) :

1. Complete the exploration of independent and weak coupled urban growth and network growth processes (all models presented in chapter 5), in order to know precisely involved mechanisms when they are virtually isolated, and to obtain morphogenesis scales.
2. Go further into the exploration of toy-model of non conventional processes such as governance network growth heuristic to pave the road for a possible integration of such modules in hybrid models.
3. Build a Marius-like generic infrastructure that implement the theory in a family of models that can be declined into diverse case studies.
4. Launch it and adapt it on these case studies.

Next steps would be too hypothetical if formulated, we propose thus to proceed iteratively in our construction of knowledge and naturally update this roadmap constantly.

- La route est longue mais la voie est libre.

Quatrième partie

OPENING

A building is never used the way it was designed, that is a reality which grasping makes the difference between good and excellent architects. The effective functional use give sense to any construction. So goes it for a knowledge edifice. We shall now take a look back on what we constructed and try to take a step back. This part develops first theoretical apparels emerging from the various aspects already tackled. It then proposes to extract fundamental open questions that future research on territorial complex systems will have to tackle in the incoming decades.

8

THEORETICAL FRAMEWORK

*Your theory is crazy, but not enough
to be true.*

- NIELS BOHR

La théorie est un élément essentiel de toute construction scientifique, en particulier en Sciences Humaines pour lesquelles la définition des objets et questions de recherche sont plus ouverts mais aussi plus déterminants des directions de recherche alors prises. Nous développons dans ce chapitre un cadre théorique autonome. Il émerge naturellement des considérations thématiques du chapitre précédent, des explorations empiriques faites dans le chapitre 4 et des expériences de modélisation conduites dans le chapitre 5. Nous proposons d'abord de construire une *Théorie Géographique* qui fixera les objets étudiés et leur nature réelle (leur ontologie), ainsi que leur interrelations. Celle-ci permettra de produire des hypothèses précises qu'on cherchera à confirmer ou infirmer par la suite.

8.1 POUR UNE THÉORIE GÉOGRAPHIQUE

8.1.1 Fondations

Territoires Humains en Réseau

Notre premier pilier a déjà été construit précédemment lors de l'exploration thématique du projet de recherche. Nous nous basons sur la notion de *Territoire Humain* élaborée par RAFFESTIN comme la base de la définition d'un système territorial. Elle permet de capturer les systèmes complexes géographiques humains dans l'ensemble de leur caractéristiques concrètes et abstraites, ainsi que dans leur représentations. Par exemple, un territoire métropolitain peut être appréhendé simplement par l'étendue fonctionnelle des flux pendulaires journaliers, ou par l'espace perçu ou vécu des différentes populations, le choix dépendant de la question précise à laquelle on cherche à répondre. Cette approche au territoire est bien sûr un choix délibéré et que d'autres entrées, possiblement compatibles, peuvent bien sûr être prises [183]. Le ciment de ce pilier est renforcé par la théorie territoriale des réseaux de DUPUY, fournissant la notion de territoire humain en réseau, comme un territoire humain dans lequel un ensemble de réseaux transactionnels potentiels ont été réalisés, ce qui s'accorde par ailleurs avec les visions du territoire comme un lieu des réseaux [56]. Nous ferons pour cela l'hypothèse fondamentale que les réseaux réels sont des éléments nécessaires des systèmes territoriaux.

Théorie Evolutive des Villes

Le second pilier de notre construction théorique est la théorie évolutive des villes de PUMAIN, en relation étroite avec l'approche complexe que nous prenons de manière générale. Cette théorie a été introduite initialement dans [202] qui argumente pour une vision dynamique des systèmes de ville, au sein desquels l'auto-organisation est essentielle. Les villes sont des entités spatiales évolutives interdépendantes dont les interrelations font émerger le comportement macroscopique à l'échelle du système de villes. Le système de villes est aussi vu comme un réseau de villes, ce qui renforce sa vision en tant que système complexe. Chaque ville est elle-même un système complexe dans l'esprit de [32], l'aspect multi-scalaire étant essentiel dans cette théorie, puisque les agents microscopiques véhiculent les processus d'évolution du système à travers des rétroactions complexes entre les échelles. Le positionnement de cette théorie au regard des Sciences des Systèmes Complexes a plus tard été confirmé [203]. Il a été montré que la théorie évolutive fournit une interprétation des lois d'échelle qui sont omniprésentes dans les systèmes urbains, qui découleraient de

la diffusion des cycles d'innovation entre les villes [209]. La notion de résilience d'un système de villes, induit par le caractère adaptatif des ces systèmes complexes, implique que les villes sont les moteurs et les adaptateurs du changement social [205]. Enfin, la dépendance au chemin est source de non-ergodicité au sein de ces systèmes, rendant les interprétations "universelles" des lois d'échelle développées par les physiciens incompatibles avec la théorie évolutive [205]. La Théorie Evolutive des Villes a été élaborée conjointement avec des modèles de systèmes urbains : par exemple le modèle Simpop2 est un modèle basé agent qui prend en compte des processus économiques, et simule sur de longues échelles de temps les motifs de croissance urbaine pour l'Europe et les Etats-unis [49]. L'accomplissement le plus récent de la théorie évolutive réside dans la production de l'ERC GeoDiversity, présentée dans [208], qui inclut à la fois des progrès techniques avancés (logiciel OpenMole), thématiques (connaissances issues des modèles SimpopLocal et Marius) et méthodologiques (modélisation par incrément). Nous interpréterons les systèmes territoriaux à la lumière de cette idée des villes comme systèmes complexes adaptatifs.

Morphogenèse

Urbaine

La notion de morphogenèse a été particulièrement soulignée par TURING dans [249] lorsqu'il proposait d'isoler des règles chimiques élémentaires qui pourraient mener à l'émergence de l'embryon et à sa forme. La morphogenèse d'un système consiste en des règles d'évolution auto-cohérentes qui produisent l'émergence de ses états successifs, i.e. la définition précise de l'auto-organisation, avec la propriété supplémentaire qu'une architecture émergente existe, au sens de relations entre la forme et la fonction. Les progrès vers la compréhension de la morphogenèse de l'embryon (en particulier l'isolation de processus particuliers induisant la différentiation de cellules à partir d'une unique) sont relativement récents grâce à l'application des approches complexes en biologie intégrative [85]. Dans le cas des systèmes urbains, l'idée de morphogenèse urbaine, i.e. de mécanismes auto-cohérents qui produiraient la forme urbaine, est plutôt utilisé dans les champs de l'architecture et de l'urbanisme [120] (comme e.g. la grammaire générative du "Pattern Language" d'ALEXANDER), en relation avec des théories de la forme urbaine [182]. Cette idée peut être poussée jusqu'à de très petites échelles comme celle du bâtiment [262] mais nous l'utiliserons plus à une échelle mesoscopique, en termes de changements d'usage du sol à une échelle intermédiaire des systèmes territoriaux, avec des ontologies similaires à la littérature de modélisation de la morphogenèse urbaine (par exemple [38] décrit un modèle de morphogenèse urbaine avec différentiation qualitative, tandis que

[172] donne un modèle de croissance urbaine basé sur une distribution monocentrique de la population perturbée par des bruits corrélés). La notion de morphogenèse sera importante dans notre théorie en lien avec la modularité et l'échelle. La modularité d'un système complexe consiste en sa décomposition en sous-modules relativement indépendants, et la décomposition modulaire d'un système peut être vue comme un moyen de supprimer les correlations non intrinsèques [144] (pour donner une image, penser à une diagonalisation par blocs d'un système dynamique du premier ordre). Dans le cadre de la conception et du contrôle de systèmes cyber-sociaux à grande échelle, des problèmes similaires surgissent naturellement et des techniques spécifiques sont nécessaires pour le passage à l'échelle des techniques simples de contrôle [256]. L'isolation d'un sous-système fournit une échelle caractéristique correspondante. Isoler des processus de morphogenèse possibles implique une extraction contrôlée (conditions au bord contrôlées par exemple) du système considéré, ce qui correspond à un niveau de modularité et donc à une échelle. Quand des processus auto-cohérents ne sont pas suffisants pour expliquer l'évolution d'un système (dans des variations raisonnables des conditions initiales), un changement d'échelle est nécessaire, causé par une transition de phase implicite dans la modularité. L'exemple de la croissance métropolitaine en est une très bonne illustration : la complexité des interactions au sein de la région métropolitaine sera croissante avec sa taille et la diversité des fonctions urbaines, ce qui conduit à un changement de l'échelle nécessaire pour comprendre les processus. L'émergence d'un aéroport international influencera fortement le développement local, ce qui correspondant à une intégration significative dans un système plus vaste. Les échelles caractéristiques et la nature des processus pour lesquels ces changements ont lieu peuvent être des questions précisément approchées par l'angle de la modélisation. Il est intéressant de noter qu'un système territorial dans lequel la morphogenèse prend sens peut être vu comme un *système auto-poiétique* au sens étendu de BOURGINE dans [44], comme un réseau de processus qui s'auto-reproduisent¹ en régulant leur conditions aux bords, ce qui souligne la notion de frontière sur laquelle nous allons finalement nous attarder.

[87] [159]

Co-évolution

Notre dernier pilier consiste en une clarification de la notion de *co-evolution*, sur laquelle HOLLAND apporte un éclairage pertinent à travers son approche des systèmes complexes adaptatifs (CAS) par

¹ qui ne sont toutefois pas cognitifs, ne rendant heureusement pas ces systèmes auto-organisés vivant au sens de auto-poiétique et cognitif

une théorie des CAS comme agents traitant des signaux grâce à leur frontières [132]. Dans cette théorie, les systèmes complexes adaptatifs forment des agrégats à différents niveaux hiérarchiques, qui correspondent à différents niveaux d'auto-organisation, et les frontières sont intriquées horizontalement et verticalement de manière complexe. Cette approche introduit la notion de *niche* comme un sous-système relativement indépendant au sein duquel les ressources circulent (de la même façon que des communautés dans un réseau) : de nombreuses illustrations telles les niches écologiques ou économiques peuvent être données. Les agents au sein d'une niche sont dits en *co-évolution*. La co-évolution implique ainsi de fortes interdépendances (impliquant des processus causaux circulaires) et une certaine indépendance au regard de l'extérieur de la niche. La notion est naturellement flexible puisqu'elle dépendra des ontologies, de la résolution, des seuils, etc. que l'on considère pour définir le système. Ce concept se transmet assez aisément à la théorie évolutive urbaine et correspond à la notion de co-évolution décrite par PUMAIN : des agents co-évolutifs dans un système de villes consistent en une niche et ses flots, signaux et limites et sont donc des entités co-évolutives au sens de HOLLAND. Cette notion sera importante pour nous dans la définition des sous-systèmes territoriaux et de leur couplage.

8.1.2 Synthèse : une théorie des systèmes territoriaux co-évolutifs en réseau

Nous synthétisons les différents piliers en une théorie géographique autonome des systèmes territoriaux pour lesquels les réseaux jouent un rôle central pour la co-évolution des composantes du système.

Pour les définitions des termes et les références, se référer à la section précédente. La formulation ici est voulue minimaliste.

Definition 1 - Système Territorial. *Un système territorial est un ensemble de territoires humains en réseau, c'est à dire des territoires humains au sein desquels et entre lesquels des réseaux réels existent.*

A cette étape la complexité et le caractère évolutif et dynamique des systèmes territoriaux sont impliqués par les parties pris mais pas une partie explicite de la théorie. We supposerons pour simplifier une définition discrète des dimensions temporelles, spatiales et ontologiques, sous des hypothèses de modularité et de stationnarité locale.

Proposition 1 - Echelle discrètes. *Supposant une décomposition modulaire discrète d'un système territorial, l'existence d'un ensemble discret (τ_i, x_i) d'échelles temporelles et fonctionnelles pour le système territorial est équivalent à la stationnarité temporelle locale d'une spécification par système dynamique stochastique du système.*

Proof (Sketch of). We underlie that any territorial system can be represented by random variables, what is equivalent to have well defined objects and states and use the Transfer Theorem on events of successive states. If $X = (X_j)$ is the modular decomposition, we have necessarily quasi-independence of components in the sense that $\text{Cov}[dX_j, dX_{j'}] \simeq 0$ at any time. General stationarity transitions induce modular transitions that are kept or not depending if they correspond to an effective transition within the subsystem, what provide temporal scales as characteristic times of sub-dynamics. Functional scales are the corresponding extent in the state space. ■

Cette proposition induit une représentation des dynamiques du système dans le temps. On peut noter que même en l'absence de représentation modulaire, le système dans son ensemble vérifiera la propriété. Cette définition des échelles permet d'introduire explicitement des boucles de rétroaction et ainsi l'émergence et la complexité, rendant la théorie compatible avec la théorie évolutive urbaine.

Assumption 1 - Imbrication des échelles et des sous-systèmes. Des réseaux complexes de rétroaction existent à la fois entre et à l'intérieur des échelles [27]. De plus, un emboîtement horizontal et vertical des limites ne sera généralement pas hiérarchique.

Au sein de ces imbrications de sous-systèmes nous pouvons isoler des composantes en co-évolution en utilisant la morphogenèse. La proposition suivante est une conséquence de l'équivalence entre l'indépendance d'une niche et sa morphogenèse. La morphogenèse fournit la décomposition modulaire (sous hypothèse de stationnarité locale) nécessaire pour l'existence de l'échelle, donnant des sous-systèmes minimaux indépendants de manière verticale (échelle) et horizontale (espace).

Proposition 2 - Co-évolution des composantes. Les processus morphogénétiques d'un système territorial sont une formulation équivalente de l'existence de sous-systèmes co-évolutifs.

Nous formulons finalement une hypothèse clé qui met les réseaux réels au centre des dynamiques co-évolutives, introduisant leur nécessité pour expliquer les processus dynamiques des systèmes territoriaux.

Assumption 2 - Nécessité des réseaux. L'évolution des réseaux ne peut pas être expliquée simplement par la dynamique des autres composantes territoriales et réciproquement, i.e. les sous-systèmes territoriaux co-évolutifs contiennent les réseaux réels. Ceux-ci peuvent ainsi être à l'origine de changements de régime (transitions entre régimes stationnaires) ou de bifurcations plus conséquentes dans les dynamiques de l'ensemble du système territorial.

Sur de longues échelles temporelles, une co-évolution globale a été montrée pour le systèmes ferroviaire français par [46]. A de plus petites échelles celle-ci est moins évidente (débat sur les effets structurants) mais nous supposons la présence d'effets co-évolutifs à toutes les échelles. Des exemples régionaux peuvent illustrer ce fait :

Lyon n'a pas les mêmes relations dynamiques avec Clermont qu'avec Saint-Etienne, et la connectivité de réseau a nécessairement un rôle à y jouer (parmi les effets des dynamiques intrinsèques des interactions, et de la distance par exemple). A une plus petite échelle encore, nous partons du principe que les effets sont encore moins observables, mais précisément à cause du fait que la co-évolution est plus forte et les bifurcations locales se produisent avec une plus grande amplitude et une plus grande fréquence que dans les systèmes macroscopiques où les attracteurs sont plus stables et les échelles de stationnarité plus grandes. Nous essayerons d'identifier des bifurcations ou des transitions de phase dans des modèles jouets, des modèles hybrides, et des analyses empiriques, à différentes échelles, sur différents cas d'études et avec différentes ontologies.

Une difficulté dans notre construction est l'hypothèse de stationnarité. Même si cela paraît une hypothèse raisonnable à de grandes échelles et a déjà été observé des données empiriques [229], nous devrons le vérifier dans nos études empiriques. En effet, cette question est au centre des efforts de recherche courants pour appliquer les techniques d'apprentissage profond aux systèmes géographiques : BOURGINE a récemment développé un cadre pour extraire des motifs des systèmes complexes adaptatifs². Les questions sont ensuite si les hypothèses de stationnarité peuvent être réglés par augmentation des états du système, et si des données hétérogènes et asynchrones peuvent être utilisées pour initialiser des séries temporelles assez longues pour une estimation correcte du réseau de neurones. Ces questions sont reliées à l'hypothèse de stationnarité pour la première et à la non-ergodicité pour la seconde.

² En utilisant un théorème de représentation [142], tout processus stationnaire discret est un *Modèle de Markov Caché*. Etant donné la définition d'un état causal comme $\mathbb{P}[\text{future}|A] = \mathbb{P}[\text{future}|B]$, la partition des états du système par la relation d'équivalence correspondantes permet de produire un *Réseau Récurrent* qui est suffisant pour déterminer l'état suivant du système, puisqu'il s'agit d'une fonction *déterministe* des états précédents et des états cachés [235] : $(x_{t+1}, s_{t+1}) = F[(x_t, s_t)]$. L'estimation des états cachés et de la fonction récurrente capture ainsi entièrement par apprentissage profond le comportement dynamique du système, i.e. l'information complète sur ses dynamiques et les processus internes.

8.2 UN CADRE THÉORIQUE POUR L'ETUDE DES SYSTÈMES SOCIAUX-TECHNIQUES

After having set up the thematic theoretical framework, we develop a more general framework in which the previous can enter. At an epistemological level, it is essential to frame generally our directions of research.

8.2.1 *Introduction*

<i>Contexte</i>	<i>Scientifique</i>
<p>Les malentendus structurels entre les Sciences Sociales et Humanités d'une part, et les dénommées Sciences Exactes d'autre part, loin d'être une généralité, semble toutefois avoir un impact conséquent sur la structure de la connaissance scientifique [129]. Plus particulièrement, le rôle de la théorie (et en fait la signification elle-même du terme) dans l'élaboration de la connaissance a une place complètement différente, en partie à cause de différentes <i>complexités perçues</i>³ des objets étudiés : par exemple, les constructions mathématiques et par extension la physique théorique sont <i>simples</i> au sens où elles sont généralement résolubles de manière analytique (ou au moins semi-analytique), tandis que les sujets des Sciences Sociales tels les humains ou la société (pour prendre un exemple préconçu) sont <i>complexes</i> au sens de systèmes complexes⁴, d'où un besoin accru d'une construction théorique (qui se base généralement sur l'empirique) pour identifier et définir qui sont nécessairement plus arbitraires dans la définition de leur limites, relations et processus, de par la multitude des points de vue possibles : PUMAIN suggère en effet dans [204] une nouvelle approche de la complexité qui serait profondément ancrée dans les sciences sociales et qui serait "mesurée par la diversité des disciplines nécessaires pour élaborer une notion". Ces différences de fond sont naturellement bénéfiques pour la diversité scientifique, mais les choses peuvent se corser quand les terrains d'étude se chevauchent, typiquement dans le cas de problématiques liées aux systèmes complexes comme déjà détaillé, comme l'exemple géographique des systèmes urbains a</p>	

³ Nous utilisons le terme *perçu* car la plupart des systèmes étudiés en physique peuvent être décrits comme simple alors qu'ils sont intrinsèquement complexe et finalement mal compris [151].

⁴ pour lesquels il n'existe pas de définition unifiée mais dont les champs d'application couvrent une étendue allant des neurosciences à la finance quantitative, en passant par exemple par la sociologie quantitative, la géographie quantitative, la biologie intégrative, etc. [185], et pour l'étude desquels diverses approches complémentaires peuvent être appliquées, comme les Systèmes Dynamiques, la Modélisation Basée Agent, la Théorie des Matrices Aléatoires

récemment montré [97]. La Science des Systèmes Complexes⁵ est présentée par certains comme "un nouveau type de science" [265], et serait au moins symptomatique d'un changement de paradigme des pratiques, des approches analytiques "exactes" vers des approches computationnelles et *evidence-based* [9], mais il est certain que cela permet de faire émerger, conjointement avec de nouvelles méthodologies, des nouveaux champs scientifiques au sens d'intérêts convergents de disciplines variées sur des questions transversales ou d'approches intégrées d'un champ particulier [43].

Objectifs

Dans ce contexte scientifique, l'étude de ce que nous désignons par *Systèmes socio-techniques*, que nous définissons de manière assez large comme des systèmes complexes hybrides qui incluent des agents ou objets sociaux qui interagissent avec des artefacts techniques et/ou un environnement naturel⁶, se situent précisément entre sciences sociales et sciences dures. L'exemple des systèmes urbains est parmi les meilleurs cas représentatifs, puisque même avant l'arrivée de nouvelles approches prétendant être "plus exactes" que les approches des sciences sociales (typiquement par des physiciens, voir e.g. le positionnement de [165], mais aussi par des chercheurs venant des sciences sociales comme BATTY [23]), une multitude d'aspects des systèmes urbains étaient déjà traités dans des sciences dures, comme l'hydrologie urbaine, la climatologie urbaine ou les aspects techniques des systèmes de transport, tandis que le centre de leur attention se reposait sur des sciences sociales comme la géographie, l'urbanisme, la sociologie, l'économie. D'où une place nécessaire de la théorie dans leur étude : suivant [161], l'étude des systèmes complexes en sciences sociales est une interaction entre analyse empirique, construction théorique et modélisation.

Nous proposons dans cette section de construire une théorie, ou plutôt un cadre théorique, pour faciliter certains aspects de l'étude de tels systèmes. De nombreuses théories existent déjà dans l'ensemble des champs liés à ce type de questionnement, et aussi à de plus haut niveaux d'abstraction concernant des méthodes comme e.g. la modélisation basée agent, mais il n'existe à notre connaissance pas de cadre théorique qui incluraient l'ensemble des points suivants que nous jugeons cruciaux (et qui peuvent être compris comme une base informelle de notre théorie) :

⁵ que nous appelons délibérément ainsi même si des débats existent sur le fait de considérer comme une science en elle-même ou comme une façon différente de faire de la science.

⁶ les systèmes géographiques au sens de [92] sont l'archetype de tels systèmes, mais cette définition peut couvrir d'autres types de systèmes comme un système de transport étendu, des systèmes sociaux pris dans un contexte environnemental, des systèmes industriels compliqués considérés avec leur utilisateurs, etc.

1. une définition précise et une emphase particulière sur la notion de couplage entre sous-systèmes, en particulier permettant de qualifier ou quantifier un certain niveau de couplage : dépendance, interdépendance, etc. entre composantes.
2. une précise définition de l'échelle, incluant l'échelle temporelle et l'échelle pour d'autres dimensions.
3. en conséquence des points précédents, une définition précise de ce qu'est un système.
4. la prise en compte de la notion d'émergence pour capturer les aspects multi-scalaires des systèmes.
5. une place centrale de l'ontologie dans la définition des systèmes, i.e. du sens dans le monde réel donné à ses objets⁷.
6. la prise en compte d'aspects hétérogènes du même système, qui peuvent être des composantes hétérogènes mais aussi des vues qui se croisent de manière complémentaire.

La suite de cette section est organisée de la façon suivante : nous construisons la théorie dans la sous-section suivante en restant à un niveau abstrait, et proposons une première application à la question des sous-systèmes co-évolutifs. Nous discutons ensuite le positionnement au regard de théories existantes, ainsi que les développements possibles et des applications concrètes.

8.2.2 *Construction de la Théorie*

Perspectives et Ontologies

Le point de départ pour construire la théorie est une approche épistémologique perspectiviste des systèmes introduite par GIERE [108]. Pour résumer, cette position interprète toute démarche scientifique comme une perspective, au sein de laquelle chacun poursuit certains objectifs et utilise ce qui est appelé *un modèle* pour les atteindre. Le modèle n'est alors rien de plus qu'un medium scientifique. VARENNE a développé [250] une typologie fonctionnelle des modèles qui peut être interprété comme un raffinement de cette théorie. Relâchons dans un premier temps cette précision potentielle et utilisons les perspectives comme des approximations des objets et concepts indéfinis. En effet, diverses visions du même objet (pouvant être complémentaires ou divergentes) ont la propriété de partager au moins l'objet lui-même, d'où notre proposition de définir les objets (et plus généralement les systèmes) à partir d'un

⁷ comme déjà expliqué précédemment, ce positionnement combiné à l'importance de la structure pourrait être relié au *Réalisme Structurel Ontologique* dans des approfondissements.

ensemble de perspectives sur ceux-ci, qui vérifient certaines propriétés que nous formalisons par la suite.

Une perspective est définie dans notre cas comme une *Dataflow Machine M* (qui correspond au model comme medium) au sens de [112] qui fournit un moyen adapté de représenter un modèle et d'y associer échelle de temps et données, à laquelle on associe un ontologie O au sens de [161], i.e. un ensemble d'éléments qui correspondent à une entité (qui peut être un objet, un agent, un processus, etc.) du monde réel. Nous incluons seulement ces deux aspects (le modèle et les objets représentés) de la théorie de Giere, en faisant l'hypothèse que le but et le producteur de la perspective sont en fait contenus dans l'ontologie s'ils font sens pour l'étude du système.

Definition 2 *Une perspective sur un système est donnée par une Dataflow Machine $M = (i, o, T)$ et une Ontologie associée O. Nous supposons que l'ontologie peut être décomposée de manière discrète en éléments atomiques $O = (O_j)_j$.*

Les éléments atomiques de l'ontologie peuvent être des constituants particuliers du systèmes, comme des agents ou des composantes, mais aussi des processus, interactions, états ou concepts par exemple. L'ontologie peut être vue comme la description exhaustive et rigoureuse du contenu de la perspective. L'hypothèse d'une *Dataflow Machine* implique que les entrées et sorties potentielles peuvent être quantifiées, ce qui n'est pas nécessairement restrictif aux perspectives quantitatives, puisque la plupart des approches qualitatives peuvent être traduites en variables discrètes à partir du moment où l'ensemble des possibles est connu ou supposé. Nous définissons alors le système de manière "réciproque", i.e. à partir d'un ensemble de perspectives sur ce qui constitue alors le système :

Definition 3 *Un système est un ensemble de perspectives sur un système : $S = (M_i, O_i)_{i \in I}$, où I n'est pas nécessairement fini.*

Nous désignons par $\mathcal{O} = (O_{j,i})_{j,i \in I}$ l'ensemble des elements dans les ontologies.

On peut noter qu'à ce stade de la construction, il n'existe pas nécessairement de cohérence structurelle sur ce qu'on appelle un système, puisque étant donné notre définition très large nous pourrions par exemple considérer un système comme une perspective sur un véhicule conjointement à une perspective sur un système de villes, ce qui ne fait pas raisonnablement sens. Des définitions approfondies et développements doivent permettre de se rapprocher des définitions classiques d'un système (entités en interaction, artefacts précisément définis, etc.). De la même manière, la définition d'un sous-système sera donnée plus loin. Les éléments

de l'approche déjà introduits permettent jusqu'ici de répondre aux points trois, cinq et six des recommandations.

PRÉCISION SUR L'ASPECT RÉCURSIF DE LA THÉORIE Une conséquence directe de ces définitions doit être détaillée : le fait qu'elles peuvent être appliquées de manière récursive. En effet, on peut imaginer prendre comme perspective un système dans notre sens, c'est à dire un ensemble de perspectives sur un système, et le faire à tout ordre. Si on considère un système à n'importe quel sens classique, alors le premier ordre peut être interprété comme une épistémologie du système, i.e. l'étude de perspectives sur un système. Une ensemble de perspectives sur des systèmes en relation peut sous certaines conditions être un domaine ou un champ d'étude, et donc un ensemble de perspectives sur divers systèmes l'épistémologie d'un champ. On peut proposer des analogies supplémentaires pour traduire l'idée derrière le caractère récursif de la théorie. C'est en effet crucial pour la signification et la cohérence de la théorie, notamment pour les raisons suivantes :

- Le choix des perspectives qui constituent un système est nécessairement subjectif et peut donc être compris comme une perspective en lui-même, et ainsi une perspective sur un système si l'on est en mesure de construire une ontologie générale.
- Nous utiliserons des relations entre ontologies par la suite, dont la construction est basée sur l'émergence est également subjective et vue comme perspectives.

Graphe

Ontologique

Nous proposons ensuite la structure du système en reliant les ontologies. Cette approche pourrait éventuellement être mise en perspective par rapport à un positionnement épistémologique de réalisme structurel [104] puisqu'une connaissance du monde est ici partiellement contenue dans la structure des modèles. Pour cette raison, nous faisons le choix d'appuyer le rôle de l'émergence, suivant l'intuition qu'il pourrait s'agir d'un outil pratique minimaliste pour capturer de façon raisonnable la structure d'un système complexe⁸. Nous prenons pour cet aspect le positionnement de BEDAU sur les différents types d'émergence, en particulier sa définition de l'émergence faible donnée dans [27]. Rappelons brièvement les définitions que nous utiliserons par la suite. BEDAU commence par définir les propriétés émergentes puis étend le concept aux phénomènes, entités, etc. De la même manière, notre cadre n'est pas restreints aux objets ou propriétés et inclut ainsi les

⁸ ce qui bien sûr ne peut être formulé comme une affirmation prouvable car cela dépendra de la définition d'un système, etc.

définitions généralisées comme lien entre ontologies. Nous appliquons la notion d'émergence sous les deux formes suivantes⁹ :

- *Emergence nominale* : une ontologie O' est inclue dans une autre ontologie O mais l'aspect de O qui est dit nominalement émergent en rapport à O' ne dépend pas de O' .
- *Emergence faible* : une partie d'une ontologie O peut être dérivée de manière computationnelle par agrégation et interactions entre les éléments d'une ontologie O' .

Comme développé précédemment, la présence d'émergence, et spécifiquement d'émergence faible, constitue une perspective en soi. Elle peut être conceptuelle et postulée comme un axiome dans une théorie thématique, mais aussi expérimentale si des traces d'émergence faible sont effectivement mesurées entre objets. Dans tous les cas, la relation entre ontologies doit être encodée dans une ontologie, ce qui n'était pas nécessairement introduit dans la définition initiale d'un système.

Nous faisons pour cette raison l'hypothèse suivante importante par la suite :

Assumption 3 *Un système peut être partiellement structuré par son extension avec une ontologie qui contient (pas nécessairement uniquement) des relations entre les éléments des ontologies de ses perspectives. Nous la désignons ontologie de couplage et supposons son existence par la suite. Nous postulons de plus son atomicité, i.e. si O est en relation avec O' , alors tout sous-ensemble de O , O' ne peuvent être en relation, ce qui n'est pas contraignant puisqu'une décomposition en des sous-ensembles indépendants assurera cette propriété si elle n'est pas vérifiée initialement.*

Cela nous permet d'exhiber des relations d'émergence pas seulement au sein d'une perspective elle-même, mais également entre les éléments de différentes perspectives. Nous définissons ensuite des relations de pré-ordre entre les sous-ensemble des ontologies :

Proposition 3 *Les relations binaires suivantes sont des pré-ordres sur $\mathcal{P}(O)$:*

- *Emergence (basée sur l'émergence faible)* : $O' \preceq O$ si et seulement si O émerge faiblement de O' .
- *Inclusion (basée sur l'émergence nominale)* : $O' \Subset O$ si et seulement si O émerge nominalement de O' .

⁹ la troisième forme rappelée par BÉDAU, l'émergence forte, ne sera pas utilisée, car nous avons besoin de capturer rien de plus des relations de dépendance et d'autonomie, et l'émergence faible est plus adéquate en termes de systèmes complexes, puisqu'elle n'assume pas "des pouvoirs causaux irréductibles" aux objets des échelles supérieures à un niveau donné. L'émergence nominale est utilisée pour capturer des relations d'inclusion entre les ontologies.

Avec la convention qu'il peut être admis qu'un objet émerge de lui-même, on a réflexivité (si une telle convention paraît absurde, on peut définir les relations comme $O \text{ émerge de } O'$ ou $O = O'$). La transitivité est clairement contenue dans la définition de l'émergence.

Notons que la relation d'inclusion est plus général qu'une inclusion entre ensembles, puisqu'elle traduit une inclusion "au sein" des éléments de l'ontologie. **TODO : give an example**

Definition 4 *Le graphe ontologique est construit par induction de la manière suivante :*

1. *Un graphe est construit, avec pour noeuds des éléments de $\mathcal{P}(O)$ et des liens de deux types : $E_W = \{(O, O') | O' \preccurlyeq O\}$ et $E_N = \{(O, O') | O' \Subset O\}$*
2. *Les noeuds sont réduits¹⁰ par : si $o \in O, O'$ et $(O' \preccurlyeq O \text{ ou } O' \Subset O)$ mais pas $(O \preccurlyeq O' \text{ or } O \Subset O')$, alors $O' \leftarrow O' \setminus o$*
3. *Les noeuds avec des ensemble se recouplant sont fusionnés, en gardant les liens liant des noeuds fusionnés. Cette étape assure des noeuds ne se recouplant pas. **TODO : what if only partially overlapping? this point is not clear***

Arbre	Ontologique	Minimal
-------	-------------	---------

La structure topologique du graphe, qui contient en un sens la *structure du système*, peut être réduite en un arbre minimal qui capture la structure hiérarchique essentielle pour la théorie.

Nous devons d'abord donner cohérence au système :

Definition 5 *Une partie cohérente du graphe ontologique est une composante faiblement connectée du graphe. Nous assumons pour la suite travailler sur une partie cohérente.*

La notion de système cohérent, ainsi que de sous-système ou d'échelle de temps des noeuds qui seront définies par la suite, nécessite de reconstruire des perspectives à partir des éléments ontologiques, i.e. l'opération inverse de ce qui a été fait dans notre procédure de deconstruction. **TODO : what is deconstructivism; position our approach in regard? Feyerabend as a confirmation?**

Assumption 4 *Il existe $O' \subset \mathcal{P}(O)$ tel que pour tout $O \subset O'$, il existe une Dataflow Machine M correspondante telle que la perspective correspondante est cohérente avec les éléments initiaux du système (i.e. les machine sont équivalentes sur les parties communes des ontologies). Si $\Phi : M \mapsto O$ est la correspondance initiale, nous notons cette construction réciproque étendue par $M' = \Phi^{<-1>}(O)$.*

¹⁰ la procédure de réduction vise à supprimer la redondance, gardant une entité au plus haut niveau où elle existe.

REMARQUE Cette hypothèse pourrait éventuellement être changée en une proposition prouvable, en supposant que l'ontologie de couplage correspond effectivement à une perspective de couplage, dont la composante *Dataflow Machine* est cohérente avec les entités couplées. Ainsi, le postulat de décomposition de [112] devrait permettre d'identifier des composantes de base correspondantes à chaque élément de l'ontologie, et construire ainsi la nouvelle perspective par induction. Nous trouvons toutefois ces hypothèses trop restrictives, puisque par exemple divers éléments de l'arbre ontologique peuvent être modélisés par la même machine irréductible, à l'image d'une équation différentielle aux variables agrégées. Nous préférons être moins restrictifs et postuler l'existence de la correspondance inverse sur certaines sous-ontologies, qui devraient être en pratique celles sur lesquelles le couplage peut effectivement être modélisé.

Grace à l'hypothèse ci-dessus, on peut définir le système cohérent comme l'image réciproque de la partie cohérente du graphe ontologique. Cela permet la connectivité du système qui est un pré-requis pour la construction de l'arbre.

Proposition 4 *La décomposition arborescente du graphe ontologique dans laquelle les noeuds contiennent les composantes fortement connexes est unique. L'arbre réduit, qui correspond au graphe ontologique les composantes fortement connexes ont été fusionnées et les liens gardés, est nommé Arbre Ontologique Minimal.*

Proof (esquisse) L'unicité découle de la définition univoque puisque les noeuds sont fixés comme les composantes fortement connexes. Il s'agit trivialement d'une décomposition en arbre puisque dans un graphe dirigé, les composantes fortement connexes ne se recoupent pas, d'où la cohérence de la décomposition.

Toute boucle $O \rightarrow O' \rightarrow \dots \rightarrow O$ dans le graphe ontologique suppose que tous ses éléments sont équivalents au sens de \preccurlyeq . Ces boucles d'équivalence devraient aider à définir la notion de couplage fort comme une application de la théorie (voir applications).

TODO : *different approaches to coupling / coupling to a certain degree using Kolmogorov etc : specific section or insert here ?*

L'Arbre Minimal Ontologique (MOT) est un arbre au sens non-dirigé, mais une forêt au sens dirigé. Sa topologie contient une certaine représentation des hiérarchies du système. Les sous-systèmes cohérents sont définis à partir de l'ensemble \mathcal{B} des branches de la forêt, comme $(\Phi^{<-1>}(\mathcal{B}), \mathcal{B})$. L'échelle de temps d'un noeud, et par extension d'un sous-système, est l'union des échelles de temps des machines correspondantes. Les niveaux de l'arbre sont définis à partir des noeuds racine, et les relations d'émergence entre les noeuds impliquent une inclusion verticale entre échelles de temps.

<i>Action</i>	<i>sur</i>	<i>des</i>	<i>Données</i>
De la même manière que les actions de groupes permettent de donner structure à l'utilisation d'un groupe sur un ensemble (généralement de données), nous ajoutons à la théorie l'aspect essentiel de relation à la réalité par une action des noeuds de l'arbre ontologique sur des ensembles de données.			

Echelles

Finally, we propose to define scales associated to a system. Following [176], an epistemological continuum of visions on scale is a consequence of differences between disciplines in the way we developed in the introduction.

This proposition is indeed compatible with our framework, as the construction of scales for each level of the ontological tree results in a broad variety of scales. Enfin, nous proposons de définir les échelles associées à un système. Suivant [176], un continuum épistémologique de visions sur l'échelle est une conséquence des différences propres à chaque discipline, comme nous avons développé en introduction. Cette proposition est en fait compatible avec notre cadre, puisque la construction d'échelles pour chaque niveau de l'arbre ontologique résulte en une grande variété d'échelles.

Soit (M, O) un sous-système et \mathbb{T} l'échelle de temps correspondante.

Nous proposons de définir "l'échelle thématique" (par exemple l'échelle spatiale) en supposant un théorème de représentation, i.e. qu'un aspect (aspect thématique) de la machine peut être représenté par une variable d'état dynamique $\vec{X}(t)$. Etant donné un opérateur d'échelle¹¹ $\|\cdot\|_S$ et que la variable d'état est différentiable à un certain niveau, l'échelle thématique est définie par $\|(d^k \vec{X}(t))_k\|_S$.

8.2.3 Application

Le cas particulier des systèmes géographiques

Dans [92], DURAND-DASTÈS introduit une définition des systèmes et structures géographiques, la structure étant le contenant spatial des systèmes vus comme des systèmes complexes ouverts en interaction (donné par ses éléments et leur attributs, les relations entre éléments et les entrée/sorties avec le monde extérieur). Pour un système donné, sa définition est une perspective, complétée par la structure pour avoir un système selon notre sens. Selon la manière dont les relations sont définies, cela peut être plus ou moins aisément d'extraire la structure ontologique.

¹¹ qui peut être de nature variée : étendue, étendue probabiliste, échelles spectrales, échelles de stationnarité, etc.

Modularité et sous-systèmes en co-évolution

Pour l'exemple des systèmes urbains, la théorie évolutive des villes entre dans ce cadre en utilisant notre théorie thématique développée dans la section précédente. La décomposition en sous-systèmes décorrélés fournit précisément des composantes fortement couplées comme des composantes en co-évolution. La correlation entre sous-systèmes devrait d'une certaine façon être corrélée à la distance topologique dans l'arbre. Si on définit les éléments d'un noeud avant réduction comme *éléments fortement couplés*, dans le cas d'ontologies dynamiques, cela fournit une définition de la *co-évolution* et de sous-systèmes en co-évolution, équivalente à la définition thématique.

8.2.4 Discussion

LIEN AVEC DES CADRES EXISTANTS Un lien avec le cadre de Cottineau-Chapron pour la multi-modélisation [63] pourrait être fait dans le cas où ils ajouteraient la couche bibliographique, qui correspondrait à la reconstruction des perspectives. [225] propose la notion de "couplage interdisciplinaire" qui est proche de notre notion de coupler des perspectives. Une correspondance avec les approches de Système de Systèmes (voir e.g. [169] pour un cadre récent englobant la modélisation et la description des systèmes) pourrait être également possible puisque nos perspectives sont construites comme des *Dataflow Machines*, mais avec la différence cruciale que la notion d'émergence est centrale dans notre cas.

CONTRIBUTION À L'ÉTUDE DES SYSTÈMES COMPLEXES Nous ne prétendons pas exhiber une théorie des systèmes (il faut généralement se méfier de la cybernétique, la systémique etc. qui ne peuvent pas tout modéliser), mais plutôt un cadre pour guider les questions de recherche (e.g. dans notre cas les conséquences directes sont les études d'épistémologie quantitative qui vient de la construction des systèmes comme perspectives ; les études empiriques pour construire des ontologies robustes pour les perspectives ; des études thématiques ciblées pour révéler des relations causales ou l'émergence pour la construction des réseaux ontologiques ; l'étude des couplages comme processus contenant possiblement de la co-évolution ; l'étude des échelles ; etc.). Cela peut être compris comme une meta-théorie dont l'application donne une théorie, la théorie thématique qui précède étant une implémentation aux systèmes territoriaux en réseau. Nous appuyons la notion de système socio-technique, croisant une approche des systèmes sociaux complexes (ontologies) avec une description des artefacts techniques (*Dataflow Machines*), prenant "le meilleur des deux mondes".

8.2.5 <i>Directions</i>	<i>de</i>	<i>Recherche</i>
-------------------------	-----------	------------------

We can draw from the construction of this theoretical framework a set of research directions, that give a general line on how trying to answer to research questions asked after the thematic theory construction.

1. The perspectivist approach implies a broad understanding of existing perspectives on a system, and of possibility of coupling between them ; thus an emphasis on applied epistemology, i.e. **Algorithmic Systematic Review** (exploration of the knowledge space), **Disciplines Mapping**(extraction of its structure) and **Datamining for Content Analysis**(refinement at the atomic level in scientific knowledge) that correspond to the three sections of chapter 3.
2. At a finer level of particularization, the knowledge of perspectives means **Knowledge of stylized facts**, i.e. empirical analysis of cases studies. These are the object of chapter 4.
3. The emphasis on coupled subsystems at different scales implies a deep understanding of coupling mechanisms, thus the need of methodological and technical developments : **Methods for Statistical Control**, **Methods for Model Exploration**, **Theoretical Study of Coupling**, **Multi-Modeling**, of which some are developed and other proposed in the methodological chapter 2.
4. Furthermore, the possibility of hidden elements within the ontology implies the test for causal relations and intermediate processes at the origin of emergence, thus e.g. the exploration of new paradigms such as role of governance within complex models as done in chapter ??.
5. Finally, the idea behind system structure contained within the ontological forest is a large set of coupled models for a given system : it means that a proper system definition (i.e. thematic problematization and exploration) and construction should yield to a structured family of models : parallel branches can be different implementations of the same process or various processes trying to explain the emerging ontology ; therefore the final objective of a family of models tackling the thematic question.

9

THEMATIC AND GENERAL PERSPECTIVES

9.1 DÉVELOPPEMENT

SPÉCIFIQUES

Le mode de communication scientifique actuel est loin d'être optimal et les initiatives se multiplient pour proposer des modèles alternatifs : la revue post-publication en est une, l'utilisation de systèmes de contrôle de version et de dépôts publics une autre, ou la publication éclair de pistes de recherche (Journal of Brief Ideas). Les descriptions courtes de pistes de recherche sont souvent reléguées à la discussion ou la conclusion des articles, qui s'écrivent de manière conventionnelle, souvent avec un biais pour justifier a posteriori l'intérêt de *sa nouvelle méthode* qu'il faut malheureusement vendre. On fait alors des plans sur la comète, propose des développements ayant peu de rapport, ou des domaines d'application *qui auront un impact* (lire qui sont à la mode ou qui reçoivent le plus de financements à la période de l'écriture). Ce manuscrit tombe bien évidemment partiellement sous ces critiques, et encore plus les articles qui lui sont associés.

Nous proposons dans cette section un exercice pas forcément conventionnel : proposer des idées et développements possibles, en s'efforçant de concrétiser les questions de recherche et/ou points techniques autant que possible, afin que ceux-ci ne s'apparentent pas à une bouteille à la mer.

9.1.1	<i>Epistémologie</i>	<i>Quantitative</i>
9.1.2	<i>Modèles</i>	<i>Multi-scalaires</i>
9.1.3	<i>Vers des Modèles</i>	<i>Opérationnels</i>

9.2 VERS UN PROGRAMME DE RECHERCHE

9.2.1 Pour une Géographie Intégrée Alternative

Comme déjà souligné en citant REY, les bouleversements techniques et méthodologiques qu'une discipline peut subir sont souvent accompagnés de profondes mutations épistémologiques, voire de la nature même de la discipline. Il est impossible de juger si l'état actuel des connaissances est transitoire, et s'il l'est quelle est le régime stable qui terminerait la transition s'il en existe un. La spéculation est le seul moyen de lever partiellement le voile, sachant que celle-ci sera nécessairement auto-réalisatrice : proposer des visions ou des programmes de recherche oriente les moyens et questions. L'incomplétude théorique en physique, lorsqu'il s'agit par exemple de lier relativité générale et physique quantique, c'est à dire le microscopique stochastique au macroscopique déterministe, orientent les visions du futur de la discipline qui elle-même conditionnent les actions concrètes qui dans ce domaine sont indispensables (financement du CERN ou de l'interféromètre d'ondes gravitationnelles spatial LISA). En géographie, même si les investissements techniques sont incomparables, ceux-ci existent (accès aux moyens de calcul, financement de laboratoires intégrés, etc.) et sont déterminés également par les perspectives pour la discipline. Nous proposons ici une vision et un manifeste d'une nouvelle géographie, qui est déjà en train de se faire et dont les bases sont solidement construites petit à petit. L'aventure de l'ERC Geodiversity en est l'allégorie, d'autant plus qu'elle a confirmé la plupart des directions professées par BANOS [13]. L'intégration de la théorie, de l'empirique, de la modélisation, mais aussi de la technique et de la méthode, n'a jamais été aussi creusée et renforcée que dans les divers développements du projet. Sans l'accès à la grille de calcul et aux nouveaux algorithmes d'exploration permis par OpenMole, les connaissances tirées du modèle SimpopLocal auraient été moindres, mais les développements techniques ont aussi été conduits par la demande thématique.

Nous proposons un cadre de connaissances pour les études ayant une composante quantitative, ou plus précisément se posant dans la lignée de la Géographie Théorique et Quantitative (TQG). Ce cadre tente de répondre aux contraintes suivantes : (i) transcender les frontières artificielles entre quantitatif et qualitatif; (ii) ne pas favoriser de composante particulière parmi les moyens de production de connaissance (aussi divers que l'ensemble des méthodes qualitatives et quantitatives classiques, les méthodes de modélisation, les approches théoriques, les données, les outils), mais bien le développement conjoint de chaque composante. Nous étendons le cadre de connaissances de [livet2010ontology], qui

consacre le triptyque des domaines empiriques, conceptuels et de la modélisation, en y ajoutant les domaines à part entière que sont les méthodes, les outils (qu'on peut voir comme des proto-méthodes) et les données. Les interactions entre chaque domaine sont détaillées, comme par exemple le passage des méthodes vers les outils qui consiste en l'implémentation, ou le passage de l'empirique aux méthodes comme prospection méthodologique. Toute démarche de production de connaissance, vue comme une *perspective* au sens de [108], est une combinaison complexe des six domaines, les fronts de connaissance dans chacun étant en co-évolution. Nous nommons notre cadre de connaissance *Géographie Intégrée*, pour souligner à la fois l'intégration des différents domaines mais aussi des connaissances qualitatives et quantitatives, puisque les deux se fondent dans chacun des domaines.

9.2.2 *Axes de Recherche*

NON-STATIONNARITÉ, NON-ERGODICITÉ ET DÉPENDANCE AU CHEMIN

COUPLAGE DES MODÈLES ET APPROCHES

CONSTRUIRE DES OUTILS DE VALIDATION POUR LES MODÈLES DE SIMULATION

EPISTÉMOLOGIE QUANTITATIVE ET EXPÉRIMENTALE POUR UNE INTÉGRATION EFFECTIVE Le mantra du mariage entre qualitatif et quantitatif est asséné mécaniquement par de nombreux auteurs, mais lorsqu'il s'agit de mise en application, on peut se permettre de soupçonner dans le meilleur des cas une naïveté, dans le pire des cas une hypocrisie. Quel sens à faire semblant de faire des analyses quantitatives en tartinant des pages de régression linéaires dont le R^2 ne dépasse pas 0.1 ? Quel sens à simuler à grande échelle des Gaussiennes pour en calculer la moyenne ?¹ Quel sens à faire semblant de détenir une connaissance qualitative fine pour justifier la mise en place de modèles relevant de l'usine à gaz technocratique ?²

POUR UNE SCIENCE TOTALEMENT OUVERTE La transparence et mise en disponibilité des données brutes ou au moins pré-traitées, et du code informatique produisant les sorties de simulation ou les figures, semble être plutôt l'exception que la règle en géographie.

¹ au moment de l'écriture, l'application étrange était toujours en ligne à <http://shiny.parisgeo.cnrs.fr/gibratsim/>, onglet simulation, malgré des signalements répétés

² cette remarque est partiellement une auto-critique, puisqu'il faut rappeler le caractère très peu qualitatif de notre travail

Comme l'assène BANOS qui y dédie un de ses commandements, "le modélisateur n'est pas le gardien de la vérité prouvée", et comme rappelé en chapitre 2, une reproductibilité parfaite des résultats est nécessaire pour une reconnaissance d'une quelconque valeur par la communauté scientifique, comme une théorie qui ne fournit pas de possibilité de falsification ne peut être considérée comme scientifique comme l'a introduit POPPER. Des expériences de revue pour *Cybergeo* ont confirmé à l'unanimité ce problème fondamental. Rappelons que la revue *PNAS* exige les données brutes et tableau produisant toute figure, pour prévenir tout biais de visualisation qu'il soit volontaire (ce qui est rédhibitoire et conduit à un signalement) ou non.

Les observateurs soulevant le caractère détraqué du mode actuel de publication scientifique sont nombreux. Un papier n'est pas un format compréhensible ni vraiment reproductible, et pousse au biais.

Comme me le rappelait un ami qui s'est spécialisé de manière admirable dans l'acceptation de papiers extrêmement techniques par des *top-journals* économiques, écrire de façon à être accepté est "un jeu" dont les règles sont subtiles et qu'il faut maîtriser pour faire carrière. Selon notre positionnement, un tel mode de communication est contraire à l'honnêteté et l'intégrité intellectuelle nécessaires à une science éthique et ouverte.

*Cet hiver a des airs de printemps
Des peuples ou de l'esprit, au diable l'âme.
Le vent se lève, ça faisait longtemps
Triste de s'enfermer pour quelques grammes.*

Cet avenir des airs de passé
S'il fallait juste trouver le régime,
Assassinée la complexité
Maintes perspectives se cachent en les crimes.

Pour une morphogenèse politique
Adieu le coron, ses tristes briques
Murs qui s'érigent tuent votre espérance.

Perle de la mer, sirène hante la crique
Du haut des tours s'amuser du cirque
L'hiver d'idées qui peuple la France.

CONCLUSION

*Explorer sans relâche les systèmes
géographiques...
- ARNAUD BANOS*

BIBLIOGRAPHIE

- [1] Alberto ABADIE, Alexis DIAMOND et Jens HAINMUELLER. "Synthetic control methods for comparative case studies : Estimating the effect of California's tobacco control program". In : *Journal of the American Statistical Association* 105.490 (2010).
- [2] Merwan ACHIBET, Stefan BALEV, Antoine DUTOT et Damien OLIVIER. "A Model of Road Network and Buildings Extension Co-evolution". In : *Procedia Computer Science* 32 (2014), p. 828–833.
- [3] Philippe AGHION, Peter HOWITT, Maxine BRANT-COLLETT et Cecilia GARCÍA-PEÑALOSA. *Endogenous growth theory*. MIT press, 1998.
- [4] Philippe AGHION, Ufuk AKCIGIT, Antonin BERGEAUD, Richard BLUNDELL et David HÉMOUS. *Innovation and Top Income Inequality*. 2015.
- [5] A. ALI, I. CARNEIRO, L. DUSSARPS, F. GUÉDEL, E LAMY, J. RAIMBAULT, L. VIGER, V. COHEN, T. Aw et S. SADEGHIAN. *Les Eco-quartiers lus par la mobilité : vers une évaluation intégrée*. Rapp. tech. Ecole des Ponts ParisTech, June 2014.
- [6] Alex ANAS, Richard ARNOTT et Kenneth A. SMALL. "Urban Spatial Structure". English. In : *Journal of Economic Literature* 36.3 (1998), pp. 1426–1464. ISSN : 00220515. URL : <http://www.jstor.org/stable/2564805>.
- [7] Philip W ANDERSON. "More is different". In : *Science* 177.4047 (1972), p. 393–396.
- [8] Joshua D ANGRIST, Guido W IMBENS et Donald B RUBIN. "Identification of causal effects using instrumental variables". In : *Journal of the American statistical Association* 91.434 (1996), p. 444–455.
- [9] W. Brian ARTHUR. *Complexity and the Shift in Modern Science*. Conference on Complex Systems, Tempe, Arizona. 2015.
- [10] Robert L AXTELL. "120 million agents self-organize into 6 million firms : a model of the US private sector". In : *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents et Multiagent Systems. 2016, p. 806–816.
- [11] Solène BAFFI. *PhD Thesis, Université Paris 1*.
- [12] Arnaud BANOS. "Pour des pratiques de modélisation et de simulation libérées en Géographies et SHS". In : *HDR. Université Paris 1* (2013).

- [13] Arnaud BANOS. "Knowledge Accelerator' in Geography and Social Sciences : Further and Faster, but Also Deeper and Wider". In : sous la dir. de Denise PUMAIN et Romain REUILLOU. in *Urban Dynamics et Simulation Models*. Springer, 2017.
- [14] Arnaud BANOS. "Pour des pratiques de modélisation et de simulation libérées en Géographie et SHS". In : *Thèse d'Habilitation à Diriger des Recherches, UMR CNRS 8504 Géographie-Cités, ISCPPIF* (Décembre 2013).
- [15] Arnaud BANOS et Cyrille GENRE-GRANDPIERRE. "Towards new metrics for urban road networks : Some preliminary evidence from agent-based simulations". In : *Agent-based models of geographical systems*. Springer, 2012, p. 627–641.
- [16] Albert-Laszlo BARABASI. "Linked : How everything is connected to everything else and what it means". In : *Plume Editors* (2002).
- [17] Ole E BARNDORFF-NIELSEN, Peter Reinhard HANSEN, Asger LUNDE et Neil SHEPHARD. "Multivariate realised kernels : consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading". In : *Journal of Econometrics* 162 (2011), p. 149–169.
- [18] C. BARRICO et C.H. ANTUNES. "Robustness Analysis in Multi-Objective Optimization Using a Degree of Robustness Concept". In : *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on.* 2006, p. 1887–1892. DOI : [10.1109/CEC.2006.1688537](https://doi.org/10.1109/CEC.2006.1688537).
- [19] Marc BARTHÉLEMY et Alessandro FLAMMINI. "Modeling urban street patterns". In : *Physical review letters* 100.13 (2008), p. 138702.
- [20] Marc BARTHÉLEMY et Alessandro FLAMMINI. "Co-evolution of density and topology in a simple model of city formation". In : *Networks and spatial economics* 9.3 (2009), p. 401–425.
- [21] Marc BARTHÉLEMY, Patricia BORDIN, Henri BERESTYCKI et Maurizio GRIBAUDI. "Self-organization versus top-down planning in the evolution of a city". In : *Scientific reports* 3 (2013).
- [22] Michael BATTY. *Cities and complexity : understanding cities with cellular automata, agent-based models, and fractals*. The MIT press, 2007.
- [23] Michael BATTY. *The new science of cities*. Mit Press, 2013.
- [24] Michael BATTY et Paul A LONGLEY. *Fractal cities : a geometry of form and function*. Academic Press, 1994.

- [25] Jean-Jacques BAVOUX, Francis BEAUCIRE, Laurent CHAPELON et Pierre ZEMBRI. *Géographie des transports*. Paris, 2005.
- [26] Sylvie BAZIN, Christophe BECKERICH, Corinne BLANQUART, Marie DELAPLACE et Ligdwine VANDENBOSSCHE. "Grande vitesse ferroviaire et développement économique local : une revue de la littérature". In : *Recherche Transports Sécurité* 27.3 (2011), p. 215–238.
- [27] Mark BDEAU. "Downward causation and the autonomy of weak emergence". In : *Principia : an international journal of epistemology* 6.1 (2002), p. 5–50.
- [28] Adel BELOUCHRANI, Moeness G AMIN et Karim ABED-MERAIM. "Direction finding in correlated noise fields based on joint block-diagonalization of spatio-temporal correlation matrices". In : *Signal Processing Letters, IEEE* 4.9 (1997), p. 266–268.
- [29] Lucien BENGUIGUI et Efrat BLUMENFELD-LIEBERTHAL. "A dynamic model for city size distribution beyond Zipf's law". In : *Physica A : Statistical Mechanics and its Applications* 384.2 (2007), p. 613–627.
- [30] Jonathan BENNETT. *OpenStreetMap*. Packt Publishing Ltd, 2010.
- [31] Laurence BERNE. "Ouverture et fermeture de territoire par les réseaux de transports dans trois espaces montagnards (Bugey, Bauges et Maurienne)". Thèse de doct. Université de Savoie, 2008.
- [32] Brian JL BERRY. "Cities as systems within systems of cities". In : *Papers in Regional Science* 13.1 (1964), p. 147–163.
- [33] Luís MA BETTENCOURT, José LOBO et Geoffrey B WEST. "Why are large cities faster? Universal scaling and self-similarity in urban organization and dynamics". In : *The European Physical Journal B-Condensed Matter and Complex Systems* 63.3 (2008), p. 285–293.
- [34] Steven BIRD. "NLTK : the natural language toolkit". In : *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics. 2006, p. 69–72.
- [35] Johan BOLLEN, David CRANDALL, Damion JUNK, Ying DING et Katy BÖRNER. "From funding agencies to scientific agency". In : *EMBO reports* 15.2 (2014), p. 131–133.
- [36] Verónica BOLÓN-CANEDO, Noelia SÁNCHEZ-MAROÑO et Amparo ALONSO-BETANZOS. "A review of feature selection methods on synthetic data". In : *Knowledge and information systems* 34.3 (2013), p. 483–519.

- [37] G. BONANNO, F. LILLO et R. N. MANTEGNA. "Levels of complexity in financial markets". In : *Physica A Statistical Mechanics and its Applications* 299 (oct. 2001), p. 16–27. eprint : [cond-mat/0104369](https://arxiv.org/abs/cond-mat/0104369).
- [38] Olivier BONIN, Jean-Paul HUBERT et al. "Modèle de morphogénèse urbaine : simulation d'espaces qualitativement différenciés dans le cadre du modèle de l'économie urbaine". In : *49è colloque de l'ASRDLF*. 2012.
- [39] Alain BONNAFOUS et François PLASSARD. "Les méthodologies usuelles de l'étude des effets structurants de l'offre de transport". In : *Revue économique* (1974), p. 208–232.
- [40] Alain BONNAFOUS, François PLASSARD et Didier SOUM. "La détection des effets structurants d'autoroute : Application à la Vallée du Rhône". English. In : *Revue économique* 25.2 (1974), pp. 233–256. ISSN : 00352764. URL : <http://www.jstor.org/stable/3500568>.
- [41] J. P. BOUCHAUD et M. POTTERS. "Financial Applications of Random Matrix Theory : a short review". In : *ArXiv e-prints* (oct. 2009). arXiv : [0910.1205 \[q-fin.ST\]](https://arxiv.org/abs/0910.1205).
- [42] J-P BOUCHAUD, Marc POTTERS et Martin MEYER. "Apparent multifractality in financial time series". In : *The European Physical Journal B-Condensed Matter and Complex Systems* 13.3 (2000), p. 595–599.
- [43] P. BOURGINE, D. CHAVALARIAS et AL. "French Roadmap for complex Systems 2008-2009". In : *ArXiv e-prints* (juil. 2009). arXiv : [0907.2221 \[nlin.AO\]](https://arxiv.org/abs/0907.2221).
- [44] Paul BOURGINE et John STEWART. "Autopoiesis and cognition". In : *Artificial life* 10.3 (2004), p. 327–345.
- [45] Catherine BOUTEILLER et Sybille BERJOAN. "Open data en transport urbain : quelles sont les données mises à disposition ? Quelles sont les stratégies des autorités organisatrices ?" In : (2013).
- [46] Anne BRETAGNOLLE. "Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis". Français. HDR. Université Panthéon-Sorbonne - Paris I, juin 2009. URL : <http://tel.archives-ouvertes.fr/tel-00459720>.
- [47] Anne BRETAGNOLLE, Fabien PAULUS et Denise PUMAIN. "Time and space scales for measuring urban growth". In : *Cybergeo : European Journal of Geography* (2002).
- [48] Anne BRETAGNOLLE et Denise PUMAIN. "Comparer deux types de systèmes de villes par la modélisation multi-agents". In : *Qu'appelle t-on aujourd'hui les sciences de la complexité ? Langages, réseaux, marchés, territoires* (2010), p. 271–299.

- [49] Anne BRETAGNOLLE et Denise PUMAIN. "Simulating Urban Networks through Multiscalar Space-Time Dynamics : Europe and the United States, 17th-20th Centuries". In : *Urban Studies* 47.13 (2010), p. 2819–2839. DOI : [10.1177/0042098010377366](https://doi.org/10.1177/0042098010377366). eprint : <http://dx.doi.org/10.1177/0042098010377366>. URL : <http://dx.doi.org/10.1177/0042098010377366>.
- [50] Anne BRETAGNOLLE, Denise PUMAIN et Céline VACCHIANI-MARCUZZO. "The organization of urban systems". In : *Complexity perspectives in innovation and social change*. Springer, 2009, p. 197–220.
- [51] Chris BRUNSDON, Stewart FOTHERINGHAM et Martin CHARLTON. "Geographically weighted regression". In : *Journal of the Royal Statistical Society : Series D (The Statistician)* 47.3 (1998), p. 431–443.
- [52] Tim Van den BULCKE, Koenraad VAN LEEMPUT, Bart NAUDTS, Piet van REMORTEL, Hongwu MA, Alain VERSCHOREN, Bart DE MOOR et Kathleen MARCHAL. "SynTReN : a generator of synthetic gene expression data for design and analysis of structure learning algorithms". In : *BMC bioinformatics* 7.1 (2006), p. 43.
- [53] Colin F CAMERER, Anna DREBER, Eskil FORSELL, Teck-Hua Ho, Jürgen HUBER, Magnus JOHANNESSEN, Michael KIRCHLER, Johan ALMENBERG, Adam ALTMEJD, Taizan CHAN et al. "Evaluating replicability of laboratory experiments in economics". In : *Science* (2016), aaf0918.
- [54] Stephen J CARVER. "Integrating multi-criteria evaluation with geographical information systems". In : *International Journal of Geographical Information System* 5.3 (1991), p. 321–339.
- [55] Junavit CHALIDABHONGSE et CC Jay KUO. "Fast motion vector estimation using multiresolution-spatio-temporal correlations". In : *Circuits and Systems for Video Technology, IEEE Transactions on* 7.3 (1997), p. 477–488.
- [56] Pierre CHAMPOILLION. "TERRITORY AND TERRITORIALIZATION : PRESENT STATE OF THE CAENTI THOUGHT". In : *International Conference of Territorial Intelligence*. INTI-International Network of Territorial Intelligence. Alba Iulia, Romania, sept. 2006, p51–58. URL : <https://halshs.archives-ouvertes.fr/halshs-00999026>.
- [57] Justin S CHANG. "Models of the Relationship between Transport and Land-use : A Review". In : *Transport Reviews* 26.3 (2006), p. 325–350.

- [58] Pierre-Olivier CHASSET, Hadrien COMMENGES, Clementine COTTINEAU et Juste RAIMBAULT. *cybergeo2020 v1.0*. Mai 2016. DOI : [10.5281/zenodo.53905](https://doi.org/10.5281/zenodo.53905). URL : <http://dx.doi.org/10.5281/zenodo.53905>.
- [59] David CHAVALARIAS et Jean-Philippe COINTET. “Phylomemetic patterns in science evolution—the rise and fall of scientific fields”. In : *Plos One* 8.2 (2013), e54847.
- [60] David CHAVALARIAS, Sylvain CHARRON, Vincent Roger DE GARDELLE et Paul BOURGINE. “Nobel, Le Jeu De La Decouverte Scientifique”. In : (2005).
- [61] Yanguang CHEN. “Urban gravity model based on cross-correlation function and Fourier analyses of spatio-temporal process”. In : *Chaos, Solitons & Fractals* 41.2 (2009), p. 603–614.
- [62] Yanguang CHEN. “Characterizing growth and form of fractal cities with allometric scaling exponents”. In : *Discrete Dynamics in Nature and Society* 2010 (2010).
- [63] Guillaume CHÉREL, Clémentine COTTINEAU et Romain REUILLOU. “Beyond Corroboration : Strengthening Model Validation by Looking for Unexpected Patterns”. In : *PLoS ONE* 10.9 (sept. 2015), e0138212. DOI : [10.1371/journal.pone.0138212](https://doi.org/10.1371/journal.pone.0138212). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0138212>.
- [64] Rémy CHICHEPORTICHE et Jean-Philippe BOUCHAUD. “A nested factor model for non-linear dependences in stock returns”. In : *arXiv preprint arXiv:1309.3102* (2013).
- [65] Paul CLAVAL. “Réseaux territoriaux et enracinement”. In : *Cahier/Groupe Réseaux* 3.7 (1987), p. 44–60.
- [66] David COLANDER. *The complexity revolution and the future of economics*. Rapp. tech. Middlebury College, Department of Economics, 2003.
- [67] H COMMENGES. “The invention of daily mobility : Performative aspects of the instruments of economics of transportation.” In : *Theses, Université Paris-Diderot-Paris VII* (2013).
- [68] Hadrien COMMENGES. “The invention of daily mobility. Performative aspects of the instruments of economics of transportation.” Theses. Université Paris-Diderot - Paris VII, déc. 2013. URL : <https://tel.archives-ouvertes.fr/tel-00923682>.
- [69] C. COTTINEAU. “MetaZipf. (Re)producing knowledge about city size distributions”. In : *ArXiv e-prints* (juin 2016). arXiv : [1606.06162 \[physics.soc-ph\]](https://arxiv.org/abs/1606.06162).

- [70] Clementine COTTINEAU. "L'évolution des villes dans l'espace post-soviétique. Observation et modélisations." Thèse de doct. Université Paris 1 Panthéon-Sorbonne, 2014.
- [71] Clémentine COTTINEAU. *Urban scaling : What cities are we talking about ?* Presentation of ongoing work at Quanturb seminar, April 1st 2015. 2015.
- [72] Clémentine COTTINEAU, Paul CHAPRON et Romain REUILLOU. "An incremental method for building and evaluating agent-based models of systems of cities". In : (2015).
- [73] Clémentine COTTINEAU, Romain REUILLOU, Paul CHAPRON, Sébastien REY-COYREHOURCQ et Denise PUMAIN. "A modular modelling framework for hypotheses testing in the simulation of urbanisation". In : *Systems* 3.4 (2015), p. 348–377.
- [74] Clémentine COTTINEAU, Florent LE NÉCHET, Marion LE TEXIER et Romain REUILLOU. "Revisiting some geography classics with spatial simulation". In : *Plurimondi. An International Forum for Research and Debate on Human Settlements*. T. 7. 15. 2015.
- [75] Thomas COURTAT, Catherine GLOAGUEN et Stephane DOUADY. "Mathematics and morphogenesis of cities : A geometrical approach". In : *Physical Review E* 83.3 (2011), p. 036106.
- [76] Noel CRESSIE et Hsin-Cheng HUANG. "Classes of nonseparable, spatio-temporal stationary covariance functions". In : *Journal of the American Statistical Association* 94.448 (1999), p. 1330–1339.
- [77] MC Cross, PC HOHENBERG et al. "Spatiotemporal chaos". In : *Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information* 263.5153 (1994), p. 1569–1569.
- [78] Yves CROZET et François DUMONT. "Retour sur les effets économiques du TGV. Les effets structurants sont un mythe (interview)". In : *Ville, Rail et Transports* 525 (2011), p. 48–51.
URL :
<https://halshs.archives-ouvertes.fr/halshs-01094554>.
- [79] Robin CURA. *Gibrat population growth simulator*. Août 2014.
DOI : [10.5281/zenodo.11415](https://doi.org/10.5281/zenodo.11415). URL :
<http://dx.doi.org/10.5281/zenodo.11415>.
- [80] Sylvain CUYALA. "Analyse spatio-temporelle d'un mouvement scientifique. L'exemple de la géographie théorique et quantitative européenne francophone." Thèse de doct. Université Paris 1 Panthéon-Sorbonne, 2014.

- [81] FD DE LEON, M FELSEN et U WILENSKY. "NetLogo Urban Suite-Tijuana Bordertowns model". In : *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL* (2007). URL : <http://ccl.northwestern.edu/netlogo/models/UrbanSuite-TijuanaBordertowns>.
- [82] Marco DE NADAI, Jacopo STAIANO, Roberto LARCHER, Nicu SEBE, Daniele QUERCIA et Bruno LEPRI. "The Death and Life of Great Italian Cities : A Mobile Phone Data Perspective". In : *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2016, p. 413–423.
- [83] Kalyanmoy DEB et Himanshu GUPTA. "Introducing robustness in multi-objective optimization". In : *Evolutionary Computation* 14.4 (2006), p. 463–494.
- [84] Guillaume DEFFUANT, Arnaud BANOS, David CHAVALARIAS, Cyrille BERTELLE, Nicolas BRODU, Pablo JENSEN, Annick LESNE, Jean-Pierre MÜLLER, Édith PERRIER et Franck VARENNE. "Visions de la complexité. Le démon de Laplace dans tous ses états". In : *Natures Sciences Sociétés* 23.1 (2015), p. 42–53.
- [85] Julien DELILE, René DOURSAT et Nadine PEYRÉAS. "Chapitre 17. Modélisation multi-agent de l'embryogenèse animale". In : *Modélisations, simulations, systèmes complexes* (2016), p. 581–624.
- [86] Jean DELONS, Nicolas COULOMBEL et Fabien LEURENT. "PIRANDELLO an integrated transport and land-use model for the Paris area". Août 2008. URL : <https://halv3-preprod.archives-ouvertes.fr/hal-00319087>.
- [87] Gaëtan DESMARAIS. "Des prémisses de la théorie de la forme urbaine au parcours morphogénétique de l'établissement humain". In : *Cahiers de géographie du Québec* 36.98 (1992), p. 251–273.
- [88] Josef DICK et Friedrich PILLICHSHAMMER. *Digital nets and sequences : Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.
- [89] Denis DIDEROT. *Entretien entre d'Alembert et Diderot*. Garnier-Flammarion, 1965.
- [90] Lynn DIRK. "A Measure of Originality The Elements of Science". In : *Social Studies of Science* 29.5 (1999), p. 765–776.
- [91] Melissa J DOBBIE et David DAIL. "Robustness and sensitivity of weighting and aggregation in constructing composite indices". In : *Ecological Indicators* 29 (2013), p. 270–277.

- [92] O DOLLFUS et F Durand DASTÈS. "Some remarks on the notions of 'structure' and 'system' in geography". In : *Geoforum* 6.2 (1975), p. 83–94.
- [93] SS DRAGOMIR. "The Ostrowski's integral inequality for Lipschitzian mappings and applications". In : *Computers & Mathematics with Applications* 38.11 (1999), p. 33–37.
- [94] Chris DRUMMOND. "Replicability is not reproducibility : nor is it good science". In : (2009).
- [95] César DUCRUET et Laurent BEAUGUITTE. "Spatial science and network science : Review and outcomes of a complex relationship". In : *Networks and Spatial Economics* 14.3-4 (2014), p. 297–316.
- [96] Gabriel DUPUY. "Vers une théorie territoriale des réseaux : une application au transport urbain". In : *Annales de Géographie*. JSTOR. 1987, p. 658–679.
- [97] Gabriel DUPUY et Lucien Gilles BENGUIGUI. "Sciences urbaines : interdisciplinarités passive, naïve, transitive, offensive". In : *Métropoles* 16 (2015).
- [98] Gilles DURANTON. "Distance, land, and proximity : economic analysis and the evolution of cities". In : *Environment and Planning a* 31.12 (1999), p. 2169–2188.
- [99] EUROSTAT. *Eurostat Geographical Data*. 2014. URL : <http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units>.
- [100] J Doyne FARMER et Duncan FOLEY. "The economy needs agent-based modelling". In : *Nature* 460.7256 (2009), p. 685–686.
- [101] Jean-Marc FAVARO et Denise PUMAIN. "Gibrat Revisited : An Urban Growth Model Incorporating Spatial Interaction and Innovation Cycles." In : *Geographical Analysis* 43.3 (2011), p. 261–286.
- [102] Jessica FRANCO, Delphine DUPUY, Olivier ROUSTANT et Astrid JOURDAN. "DiceDesign-package". In : *Designs of Computer Experiments* (2009), p. 2.
- [103] Morgan R FRANK, Jake Ryland WILLIAMS, Lewis MITCHELL, James P BAGROW, Peter Sheridan DODDS et Christopher M DANFORTH. "Constructing a taxonomy of fine-grained human movement and activity motifs through social media". In : *arXiv preprint arXiv:1410.1393* (2014).
- [104] Roman FRIGG et Ioannis VOTSI. "Everything you always wanted to know about structural realism but were afraid to ask". In : *European journal for philosophy of science* 1.2 (2011), p. 227–276.

- [105] Xavier GABAIX. "Zipf's law for cities : an explanation". In : *Quarterly journal of Economics* (1999), p. 739–767.
- [106] Xavier GABAIX et Yannis M. IOANNIDES. "Chapter 53 The evolution of city size distributions". In : *Cities and Geography*. Sous la dir. de J. Vernon HENDERSON et Jacques-François THISSE. T. 4. *Handbook of Regional and Urban Economics*. Elsevier, 2004, p. 2341 –2378. DOI : [http://dx.doi.org/10.1016/S1574-0080\(04\)80010-5](http://dx.doi.org/10.1016/S1574-0080(04)80010-5). URL : <http://www.sciencedirect.com/science/article/pii/S1574008004800105>.
- [107] Murray GELL-MANN. *The Quark and the Jaguar : Adventures in the Simple and the Complex*. Macmillan, 1995.
- [108] Ronald N GIERE. *Scientific perspectivism*. University of Chicago Press, 2010.
- [109] Frédéric GILLI et Jean-Marc OFFNER. *Paris, métropole hors les murs : aménager et gouverner un Grand Paris*. Sciences Po, les presses, 2009.
- [110] Jean-François GIRRES et Guillaume TOUYA. "Quality assessment of the French OpenStreetMap dataset". In : *Transactions in GIS* 14.4 (2010), p. 435–459.
- [111] Jean-François GLEYZE. "La vulnérabilité structurelle des réseaux de transport dans un contexte de risques". Thèse de doct. Université Paris-Diderot-Paris VII, 2005.
- [112] Boris GOLDEN, Marc AIGUER et Daniel KROB. "Modeling of complex systems ii : A minimalist and unified semantics for heterogeneous integrated systems". In : *Applied Mathematics and Computation* 218.16 (2012), p. 8039–8055.
- [113] Isaac GOLDHIRSCH, Pierre-Louis SULEM et Steven A ORSZAG. "Stability and Lyapunov stability of dynamical systems : A differential approach and a numerical method". In : *Physica D : Nonlinear Phenomena* 27.3 (1987), p. 311–337.
- [114] Daniel A GRIFFITH. "Towards a theory of spatial statistics". In : *Geographical Analysis* 12.4 (1980), p. 325–339.
- [115] Daniel A GRIFFITH. "What is spatial autocorrelation ? Reflections on the past 25 years of spatial statistics". In : *Espace géographique* 21.3 (1992), p. 265–280.
- [116] Daniel A GRIFFITH. *Advanced spatial statistics : special topics in the exploration of quantitative spatial data series*. T. 12. Springer Science & Business Media, 2012.

- [117] Volker GRIMM, Eloy REVILLA, Uta BERGER, Florian JELTSCH, Wolf M MOOIJ, Steven F RAILSBACK, Hans-Hermann THULKE, Jacob WEINER, Thorsten WIEGAND et Donald L DEANGELIS. "Pattern-oriented modeling of agent-based complex systems : lessons from ecology". In : *science* 310.5750 (2005), p. 987–991.
- [118] Marianne GUÉROIS et Renaud LE GOIX. "La dynamique spatio-temporelle des prix immobiliers à différentes échelles : le cas des appartements anciens à Paris (1990-2003)". In : *Cybergeo : European Journal of Geography* (2009).
- [119] Xiaolei GUO et Henry X LIU. "Bounded rationality and irreversible network change". In : *Transportation Research Part B : Methodological* 45.10 (2011), p. 1606–1618.
- [120] R. HACHI. *La fractalité comme indicateur de l'état de conservation du patrimoine urbain, Master Thesis Memoire*. Rapp. tech. Université Paris VII, 2013.
- [121] Ian HACKING. *The social construction of what?* Harvard university press, 1999.
- [122] Herman HAKEN et Juval PORTUGALI. "The face of the city is its information". In : *Journal of Environmental Psychology* 23.4 (2003), p. 385–408.
- [123] Hermann HAKEN. "Synergetics". In : *Naturwissenschaften* 67.3 (1980), p. 121–128.
- [124] Mordechai HAKLAY. "How good is volunteered geographical information ? A comparative study of OpenStreetMap and Ordnance Survey datasets". In : *Environment and planning B : Planning and design* 37.4 (2010), p. 682–703.
- [125] Sangjin HAN. "Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks". In : *Transportation Research Part B : Methodological* 37.3 (2003), p. 225–249.
- [126] Michael S HANSEN, Christof BALTES, Jeffrey TSAO, Sebastian KOZERKE, Klaas P PRUESSMANN, Peter BOESIGER et Erik M PEDERSEN. "Accelerated dynamic Fourier velocity encoding by exploiting velocity-spatio-temporal correlations". In : *Magnetic Resonance Materials in Physics, Biology and Medicine* 17.2 (2004), p. 86–94.
- [127] EGP HARAN et Daniel R Vining. "A MODIFIED YULE-SIMON MODEL ALLOWING FOR INTERCITY MIGRATION AND ACCOUNTING FOR THE OBSERVED FORM OF THE SIZE DISTRIBUTION OF CITIES*". In : *Journal of Regional Science* 13.3 (1973), p. 421–437.

- [128] A HERNANDO, R HERNANDO, A PLASTINO et E ZAMBRANO. "Memory-endowed US cities and their demographic interactions". In : *Journal of The Royal Society Interface* 12.102 (2015), p. 20141185.
- [129] C. A. HIDALGO. "Disconnected ! The parallel streams of network literature in the natural and social sciences". In : *ArXiv e-prints* (nov. 2015). arXiv : [1511.03981 \[physics.soc-ph\]](https://arxiv.org/abs/1511.03981).
- [130] Bill HILLIER. "The Fourth Sustainability, Creativity : Statistical Associations and Credible Mechanisms". In : *Complexity, Cognition, Urban Planning and Design*. Springer, 2016, p. 75–92.
- [131] Bill HILLIER et Julienne HANSON. *The social logic of space*. Cambridge university press, 1989.
- [132] John H HOLLAND. *Signals and boundaries : Building blocks for complex adaptive systems*. Mit Press, 2012.
- [133] C. HOLMES, M. GHAFARI, A. ANZAR, V. SARAVANAN et I. NEMENMAN. "Luria-Delbrück, revisited : The classic experiment does not rule out Lamarckian evolution". In : *ArXiv e-prints* (jan. 2017). arXiv : [1701.05627 \[q-bio.PE\]](https://arxiv.org/abs/1701.05627).
- [134] Marius HOMOCIANU. "Transport-land use interaction modeling - Residential choices of households in urban area of Lyon". Theses. Université Lumière - Lyon II, jan. 2009. URL : <https://tel.archives-ouvertes.fr/tel-00359302>.
- [135] *Hypergeo*. URL : <http://www.hypergeo.eu/spip.php?page=sommaire>.
- [136] Michael IACONO, David LEVINSON et Ahmed EL-GENEIDY. "Models of transportation and land use change : a guide to the territory". In : *Journal of Planning Literature* 22.4 (2008), p. 323–340.
- [137] Robert A JARROW. "In Honor of the Nobel Laureates Robert C. Merton and Myron S. Scholes : A Partial Differential Equation that Changed the World". In : *The Journal of Economic Perspectives* (1999), p. 229–248.
- [138] Anne JÉGOU, Vincent AUGISEAU, Cécile GUYOT, Cécile JUDÉAUX, François-Xavier MONACO, Pierre PECH et al. "L'évaluation par indicateurs : un outil nécessaire d'aménagement urbain durable ?. Réflexions à partir de la démarche parisienne pour le géographe et l'aménageur". In : *Cybergeo : European Journal of Geography* (2012).
- [139] Alexandros KARATZOGLOU, Alex SMOLA, Kurt HORNIK et Achim ZEILEIS. "kernlab – An S4 Package for Kernel Methods in R". In : *Journal of Statistical Software* 11.9 (2004), p. 1–20. URL : <http://www.jstatsoft.org/v11/i09/>.

- [140] Yan KE, Rahul SUKTHANKAR et Martial HEBERT. "Spatio-temporal shape and flow correlation for action recognition". In : *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.* IEEE. 2007, p. 1–8.
- [141] Marie-Laurence KEERSMAECKER, Pierre FRANKHAUSER et Isabelle THOMAS. "Using fractal dimensions for characterizing intra-urban diversity : The example of Brussels". In : *Geographical analysis* 35.4 (2003), p. 310–328.
- [142] Frank B KNIGHT. "A predictive view of continuous time processes". In : *The annals of Probability* (1975), p. 573–596.
- [143] Christof KOCH et Gilles LAURENT. "Complexity and the nervous system". In : *Science* 284.5411 (1999), p. 96–98.
- [144] A. KOLCHINSKY, A. J. GATES et L. M. ROCHA. "Modularity and the Spread of Perturbations in Complex Dynamical Systems". In : *ArXiv e-prints* (sept. 2015). arXiv : [1509.04386 \[physics.soc-ph\]](https://arxiv.org/abs/1509.04386).
- [145] Marko KRYVOBOKOV, Jean-Baptiste CHESNEAU, Alain BONNAFOUS, Jean DELONS et Vincent PIRON. "Comparison of Static and Dynamic Land Use-Transport Interaction Models". In : *Transportation Research Record : Journal of the Transportation Research Board* 2344.1 (2013), p. 49–58.
- [146] Mei-Po KWAN. "Space-time and integral measures of individual accessibility : a comparative analysis using a point-based framework". In : *Geographical analysis* 30.3 (1998), p. 191–216.
- [147] LA POSITION DE LA REVUE SOCIÉTÉS DANS L'ESPACE DISCURSIF DE LA SOCIOLOGIE FRANÇAISE. URL : <http://zilsel.hypotheses.org/category/canular>.
- [148] L'ESPACE GÉOGRAPHIQUE. *Les effets structurants des infrastructures de transport, L'Espace géographique 2014/1 (Tome 43)*, p. 51–67. 2014.
- [149] Alain L'HOSTIS, Claude SOULAS, Gebhard WULFHORST et al. "La ville orientée vers le rail". In : *Ville et mobilité* (2012).
- [150] C. LAGESSE. "Read Cities through their Lines. Methodology to characterize spatial graphs". In : *ArXiv e-prints* (déc. 2015). arXiv : [1512.01268 \[physics.soc-ph\]](https://arxiv.org/abs/1512.01268).
- [151] Robert B LAUGHLIN. *A different universe : Reinventing physics from the bottom down*. Basic Books, 2006.
- [152] Robert L LAUNER et Graham N WILKINSON. *Robustness in statistics*. Academic Press, 2014.

- [153] Florent LE NÉCHET. "De la forme urbaine à la structure métropolitaine : une typologie de la configuration interne des densités pour les principales métropoles européennes de l'Audit Urbain". In : *Cybergeo : European Journal of Geography* (2015).
- [154] Thomas LECHNER, Ben WATSON, Pin REN, Uri WILENSKY, Seth TISUE et Martin FELSEN. "Procedural modeling of land use in cities". In : (2004).
- [155] Fabien LEURENT et Houda BOUJNAH. "A user equilibrium, traffic assignment model of network route and parking lot choice, with search circuits and cruising flows". In : *Transportation Research Part C : Emerging Technologies* 47 (2014), p. 28–46.
- [156] David Matthew LEVINSON, Feng XIE et Shanjiang ZHU. "The co-evolution of land use and road networks". In : *Transportation and traffic theory* (2007), p. 839–859.
- [157] David LEVINSON. "Density and dispersion : the co-development of land use and rail in London". In : *Journal of Economic Geography* 8.1 (2008), p. 55–77.
- [158] David LEVINSON, Wei CHEN et al. "Paving new ground". In : () .
- [159] Albert Lévy. "Formes urbaines et significations : revisiter la morphologie urbaine". In : *Espaces et sociétés* 3 (2005), p. 25–48.
- [160] Michael LISSACK. "Subliminal influence or plagiarism by negligence ? The Slodderwetenschap of ignoring the internet". In : *Journal of Academic Ethics* (2013).
- [161] Pierre LIVET, Jean-Pierre MULLER, Denis PHAN et Lena SANDERS. "Ontology, a Mediator for Agent-Based Modeling in Social Science". In : *Journal of Artificial Societies and Social Simulation* 13.1 (2010), p. 3. ISSN : 1460-7425. URL : <http://jasss.soc.surrey.ac.uk/13/1/3.html>.
- [162] Thomas LOUAIL, Maxime LENORMAND, Juan Murillo ARIAS et José J RAMASCO. "Crowdsourcing the Robin Hood effect in cities". In : *arXiv preprint arXiv :1604.08394* (2016).
- [163] R. LOUF et M. BARTHELEMY. "How congestion shapes cities : from mobility patterns to scaling". In : *ArXiv e-prints* (jan. 2014). arXiv : [1401.8200 \[physics.soc-ph\]](https://arxiv.org/abs/1401.8200).
- [164] Rémi LOUF et Marc BARTHELEMY. "A typology of street patterns". In : *Journal of The Royal Society Interface* 11.101 (2014), p. 20140924.
- [165] Rémi LOUF et Marc BARTHELEMY. "Scaling : lost in the smog". In : *arXiv preprint arXiv :1410.4964* (2014).

- [166] Rémi LOUF et Marc BARTHELEMY. "Patterns of residential segregation". In : *arXiv preprint arXiv :1511.04268* (2015).
- [167] Rémi LOUF, Pablo JENSEN et Marc BARTHELEMY. "Emergence of hierarchy in cost-driven growth of spatial networks". In : *Proceedings of the National Academy of Sciences* 110.22 (2013), p. 8824–8829.
- [168] Qiang LUO, Wenlian LU, Wei CHENG, Pedro A VALDES-SOSA, Xiaotong WEN, Mingzhou DING et Jianfeng FENG. "Spatio-temporal Granger causality : A new framework". In : *NeuroImage* 79 (2013), p. 241–263.
- [169] Dominique LUZEAUX. "A formal foundation of systems engineering". In : *Complex Systems Design & Management*. Springer, 2015, p. 133–148.
- [170] Chunsheng MA. "Spatio-temporal covariance functions generated by mixtures". In : *Mathematical geology* 34.8 (2002), p. 965–975.
- [171] Hani S MAHMASSANI et Gang-Len CHANG. "On boundedly rational user equilibrium in transportation systems". In : *Transportation science* 21.2 (1987), p. 89–99.
- [172] Hernán A MAKSE, José S ANDRADE, Michael BATTY, Shlomo HAVLIN, H Eugene STANLEY et al. "Modeling urban growth patterns with correlated percolation". In : *Physical Review E* 58.6 (1998), p. 7054.
- [173] David MANGIN. *Le Grand Paris, où en est-on ?* Conférence de David Mangin le 14 mars 2014, ENPC et ENSAVT. 2014.
- [174] David MANGIN et Philippe PANERAI. *Projet urbain*. Parenthèses, 1999.
- [175] Steven M MANSON. "Simplifying complexity : a review of complexity theory". In : *Geoforum* 32.3 (2001), p. 405–414.
- [176] Steven M MANSON. "Does scale exist ? An epistemological scale continuum for complex human–environment systems". In : *Geoforum* 39.2 (2008), p. 776–788.
- [177] Rosario Nunzio MANTEGNA, Harry Eugene STANLEY et al. *An introduction to econophysics : correlations and complexity in finance*. T. 9. Cambridge university press Cambridge, 2000.
- [178] Caterina MARCHIONNI. "Geographical economics versus economic geography : towards a clarification of the dispute". In : *Environment and Planning A* 36.10 (2004), p. 1737–1753.
- [179] R Timothy MARLER et Jasbir S ARORA. "Survey of multi-objective optimization methods for engineering". In : *Structural and multidisciplinary optimization* 26.6 (2004), p. 369–395.

- [180] MENDELEY. *Mendeley Reference Manager*.
<http://www.mendeley.com/>. 2015.
- [181] Rolf MOECKEL, Klaus SPIEKERMANN et Michael WEGENER. "Creating a synthetic population". In : *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)*. 2003.
- [182] Anne Vernez MOUDON. "Urban morphology as an emerging interdisciplinary field". In : *Urban morphology* 1.1 (1997), p. 3–10.
- [183] Alexander B MURPHY. "Entente territorial : Sack and Raffestin on territoriality". In : *Environment and Planning D : Society and Space* 30.1 (2012), p. 159–172.
- [184] M. E. J. NEWMAN. "Prediction of highly cited papers". In : *ArXiv e-prints* (oct. 2013). arXiv : [1310.8220](https://arxiv.org/abs/1310.8220) [physics.soc-ph].
- [185] MEJ NEWMAN. "Complex systems : A survey". In : *arXiv preprint arXiv:1112.1440* (2011).
- [186] Mark EJ NEWMAN. "The structure and function of complex networks". In : *SIAM review* 45.2 (2003), p. 167–256.
- [187] H NIEDERREITER. "Discrepancy and convex programming". In : *Annali di matematica pura ed applicata* 93.1 (1972), p. 89–97.
- [188] David O'SULLIVAN et Steven M MANSON. "Do Physicists Have 'Geography Envy'? And What Can Geographers Learn From It?" In : *Annals of the Association of American Geographers* (2015).
- [189] Oliver O'BRIEN, James CHESHIRE et Michael BATTY. "Mining bicycle sharing data for generating insights into sustainable transport systems". In : *Journal of Transport Geography* 34 (2014), p. 262–273.
- [190] Jean-Marc OFFNER. "Les "effets structurants" du transport : mythe politique, mystification scientifique". In : *Espace géographique* 22.3 (1993), p. 233–242.
- [191] Jean-Marc OFFNER et Denise PUMAIN. "Réseaux et territoires-significations croisées". In : (1996).
- [192] OPENSTREETMAP. *OpenStreetMap*. 2012.
- [193] S. & al. OSTROWETSKY. "Les Villes Nouvelles, 30 ans après". In : *Espaces et Sociétés* n°119, 4/2004 (2004).
- [194] Neil J PAULLEY et F Vernon WEBSTER. "Overview of an international study to compare models and evaluate land-use and transport policies". In : *Transport Reviews* 11.3 (1991), p. 197–222.

- [195] Antoine PICON. *Smart cities : théorie et critique d'un idéal auto-réalisateur*. B2, 2013.
- [196] Bruce Wm PIGOZZI. "Interurban linkages through polynomially constrained distributed lags". In : *Geographical Analysis* 12.4 (1980), p. 340–352.
- [197] Thomas PIKETTY. *Le capital au XXIe siècle*. Seuil, 2013.
- [198] Y. POTIRON. "Estimating the integrated parameter of the locally parametric model in high-frequency data." In : *Working Paper* (2016).
- [199] Yoann POTIRON et Per MYKLAND. "Estimation of integrated quadratic covariation between two assets with endogenous sampling times". In : *arXiv preprint arXiv :1507.01033* (2015).
- [200] Ilya PRIGOGINE et Isabelle STENGERS. *The end of certainty*. Simon et Schuster, 1997.
- [201] David R PRITCHARD et Eric J MILLER. "Advances in agent population synthesis and application in an integrated land use and transportation model". In : *Transportation Research Board 88th Annual Meeting*. 09-1686. 2009.
- [202] Denise PUMAIN. "Pour une théorie évolutive des villes". In : *Espace géographique* 26.2 (1997), p. 119–134.
- [203] Denise PUMAIN. "Une approche de la complexité en géographie". In : *Géocarrefour* 78.1 (2003), p. 25–31.
- [204] Denise PUMAIN. "Cumulativité des connaissances". In : *Revue européenne des sciences sociales. European Journal of Social Sciences* XLIII-131 (2005), p. 5–12.
- [205] Denise PUMAIN. "Une théorie géographique des villes". In : *Bulletin de la Société géographie de Liège* 55 (2010), p. 5–15.
- [206] Denise PUMAIN. "Multi-agent system modelling for urban systems : The series of SIMPOP models". In : *Agent-based models of geographical systems*. Springer, 2012, p. 721–738.
- [207] Denise PUMAIN. "Urban systems dynamics, urban growth and scaling laws : The question of ergodicity". In : *Complexity Theories of Cities Have Come of Age*. Springer, 2012, p. 91–103.
- [208] Denise PUMAIN et Romain REUILLOU. *Urban Dynamics and Simulation Models*. 2017.
- [209] Denise PUMAIN, Fabien PAULUS, Céline VACCHIANI-MARCUZZO et José LOBO. "An evolutionary theory for interpreting urban scaling laws". In : *Cybergeo : European Journal of Geography* (2006).
- [210] Stephen H PUTMAN. "Urban land use and transportation models : A state-of-the-art summary". In : *Transportation Research* 9.2 (1975), p. 187–202.

- [211] Rami PUZIS, Yaniv ALTHULER, Yuval ELOVICI, Shlomo BEKHOR, Yoram SHIFTAN et Alex PENTLAND. "Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks". In : *Journal of Intelligent Transportation Systems* 17.1 (2013), p. 91–105.
- [212] DT QGis. "Quantum GIS geographic information system". In : *Open Source Geospatial Foundation Project* (2011).
- [213] R CORE TEAM. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL : <http://www.R-project.org/>.
- [214] Claude RAFFESTIN. "Repères pour une théorie de la territorialité humaine". In : (1988).
- [215] J. RAIMBAULT. "Calibration of a Spatialized Urban Growth Model". In : *Working Paper, draft at https://gitub.com/JusteRaimbault/CityNetwork/tree/master/Docs/Papers/Density* (2016).
- [216] J. RAIMBAULT. *Vers des Modèles Couplant Développement Urbain et Croissance des Réseaux de Transports, PhD Project Description*. Rapp. tech. Géographie-Cités UMR CNRS 8504/LVMT UMR-T IFSTTAR 9403, October 2014.
- [217] J. RAIMBAULT, A. BANOS et R. DOURSAT. "A hybrid network/grid model of urban morphogenesis and optimization". In : *Proceedings of the 4th International Conference on Complex Systems and Applications (ICCSA 2014), June 23-26, 2014, Université de Normandie, Le Havre, France*; M. A. Aziz-Alaoui, C. Bertelle, X. Z. Liu, D. Olivier, eds. : pp. 51-60. 2014.
- [218] J. RAIMBAULT et J. GONZALEZ. "Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d'Infrastructures de Transport". In : *Rencontres du Labex Dynamites*. May 2015.
- [219] Juste RAIMBAULT. "Towards Models Coupling Urban Growth and Transportation Network Growth. First year preliminary memoire. DOI : <http://dx.doi.org/10.5281/zenodo.60538>". Thèse de doct. Université Paris-Diderot - Paris VII, 2016.
- [220] Karthik RAM. "Git can facilitate greater reproducibility and increased transparency in science." In : *Source code for biology and medicine* 8.1 (2013), p. 7.
- [221] James B RAMSEY. "Wavelets in economics and finance : Past and future". In : *Studies in Nonlinear Dynamics & Econometrics* 6 (2002).

- [222] Thomas Kjær RASMUSSEN, David Paul WATLING, Carlo Giacomo PRATO et Otto Anker NIELSEN. "Stochastic user equilibrium with equilibrated choice sets : Part II-Solving the restricted SUE for the logit family". In : *Transportation Research Part B : Methodological* 77 (2015), p. 146–165.
- [223] Romain REUILLON, Mathieu LECLAIRE et Sébastien REY-COYREHOURCQ. "OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models". In : *Future Generation Computer Systems* 29.8 (2013), p. 1981–1990.
- [224] Sébastien REY-COYREHOURCQ. "Une plateforme intégrée pour la construction et Une plateforme intégrée pour la construction et l'évaluation de modèles de simulation en géographie". Thèse de doct. Université Paris 1 Panthéon-Sorbonne, 2015.
- [225] Henri REYMOND et Colette CAUVIN. "La logique ternaire de Stéphane Lupasco et le raisonnement géocartographique bioculturel d'*Homo geographicus*. L'apport de la notion de couplage transdisciplinaire dans l'approche de l'agrégation morphologique des agglomérations urbaines". In : *Cybergeo : European Journal of Geography* (2013).
- [226] Gerta RUCKER. "Network meta-analysis, electrical networks and graph theory". In : *Research Synthesis Methods* 3.4 (2012), p. 312–324.
- [227] Yikang RUI et Yifang BAN. "Urban growth modeling with road network expansion and land use development". In : *Advances in Cartography and GIScience. Volume 2*. Springer, 2011, p. 399–412.
- [228] Yikang RUI, Yifang BAN, Jiechen WANG et Jan HAAS. "Exploring the patterns and evolution of self-organized urban street networks through modeling". In : *The European Physical Journal B* 86.3 (2013), p. 1–8.
- [229] Lena SANDERS. *Système de villes et synergétique*. Economica, 1992.
- [230] Lena SANDERS, Denise PUMAIN, Hélène MATHIAN, France GUÉRIN-PACE et Stephane BURA. "SIMPOP : a multiagent system for the study of urbanism". In : *Environment and Planning B* 24 (1997), p. 287–306.
- [231] E. SARIGÖL, R. PFITZNER, I. SCHOLTES, A. GARAS et F. SCHWEITZER. "Predicting Scientific Success Based on Coauthorship Networks". In : *ArXiv e-prints* (fév. 2014). arXiv : [1402.7268 \[physics.soc-ph\]](https://arxiv.org/abs/1402.7268).

- [232] Thomas C SCHELLING. "Dynamic models of segregation". In : *Journal of mathematical sociology* 1.2 (1971), p. 143–186.
- [233] Clara SCHMITT. "Modélisation de la dynamique des systèmes de peuplement : de SimpopLocal à SimpopNet." Thèse de doct. Paris 1, 2014.
- [234] Clara SCHMITT, Sébastien REY-COYREHOURCQ, Romain REUILLO et Denise PUMAIN. "Half a billion simulations : Evolutionary algorithms and distributed computing for calibrating the SimpopLocal geographical model". In : (2014).
- [235] Cosma Rohilla SHALIZI et James P CRUTCHFIELD. "Computational mechanics : Pattern and prediction, structure and simplicity". In : *Journal of statistical physics* 104.3-4 (2001), p. 817–879.
- [236] Herbert A. SIMON. "On a Class of Skew Distribution Functions". English. In : *Biometrika* 42.3/4 (1955), pp. 425–440. ISSN : 00063444. URL : <http://www.jstor.org/stable/2333389>.
- [237] Taoufik SOUAMI. *Ecoquartiers : secrets de fabrication*. Scrineo, 2012.
- [238] H Eugene STANLEY, Luis A Nunes AMARAL, David CANNING, Parameswaran GOPIKRISHNAN, Youngki LEE et Yanhui LIU. "Econophysics : Can physicists contribute to the science of economics?" In : *Physica A : Statistical Mechanics and its Applications* 269.1 (1999), p. 156–169.
- [239] Forrest R. STEVENS, Andrea E. GAUGHAN, Catherine LINARD et Andrew J. TATEM. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data". In : *PLoS ONE* 10.2 (fév. 2015), p. 1–22. DOI : [10.1371/journal.pone.0107042](https://doi.org/10.1371/journal.pone.0107042). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0107042>.
- [240] Victoria STODDEN. "The scientific method in practice : Reproducibility in the computational sciences". In : (2010).
- [241] JL SULLIVAN, DC NOVAK, L AULTMAN-HALL et David M SCOTT. "Identifying critical road segments and measuring system-wide robustness in transportation networks with isolating links : a link-based capacity-reduction approach". In : *Transportation Research Part A : Policy and Practice* 44.5 (2010), p. 323–336.
- [242] R Core TEAM. *R Language Definition*. 2000.

- [243] Atsushi TERO, Seiji TAKAGI, Tetsu SAIGUSA, Kentaro Ito, Dan P. BEBBER, Mark D. FRICKER, Kenji YUMIKI, Ryo KOBAYASHI et Toshiyuki NAKAGAKI. "Rules for Biologically Inspired Adaptive Network Design". In : *Science* 327.5964 (2010), p. 439–442. DOI : [10.1126/science.1177894](https://doi.org/10.1126/science.1177894). eprint : <http://www.sciencemag.org/content/327/5964/439.full.pdf>. URL : <http://www.sciencemag.org/content/327/5964/439.abstract>.
- [244] Mihai TIVADAR, Yves SCHAEFFER, André TORRE et Frédéric BRAY. "OASIS—un Outil d'Analyse de la Ségrégation et des Inégalités Spatiales". In : *Cybergeo : European Journal of Geography* (2014).
- [245] Antoine TORDEUX et Sylvain LASSARRE. "Jam avoidance with autonomous systems". In : *arXiv preprint arXiv :1601.07713* (2016).
- [246] Yu-Hsin TSAI. "Quantifying urban form : compactness versus' sprawl'". In : *Urban studies* 42.1 (2005), p. 141–161.
- [247] Ruey S. TSAY. *MTS : All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*. R package version 0.33. 2015. URL : <http://CRAN.R-project.org/package=MTS>.
- [248] Michele TUMMINELLO, Tomaso ASTE, Tiziana Di MATTEO et Rosario N MANTEGNA. "A tool for filtering information in complex systems". In : *Proceedings of the National Academy of Sciences of the United States of America* 102 (2005), p. 10421–10426.
- [249] Alan Mathison TURING. "The chemical basis of morphogenesis". In : *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 237.641 (1952), p. 37–72.
- [250] Franck VARENNE. "Framework for M&S with Agents in Regard to Agent Simulations in Social Sciences". In : *Activity-Based Modeling and Simulation* (2010), p. 53–84.
- [251] Franck VARENNE. "Les simulations computationnelles dans les sciences sociales". In : *Nouvelles Perspectives en Sciences Sociales* 5.2 (2010), p. 17–49.
- [252] Franck VARENNE, Marc SILBERSTEIN et al. *Modéliser & simuler. Epistémologies et pratiques de la modélisation et de la simulation, tome 1*. 2013.
- [253] Suzanne VARET. "Développement de méthodes statistiques pour la prédiction d'un gabarit de signature infrarouge". Thèse de doct. Université Paul Sabatier-Toulouse III, 2010.

- [254] Mehmet C VURAN, Özgür B AKAN et Ian F AKYILDIZ. "Spatio-temporal correlation : theory and applications for wireless sensor networks". In : *Computer Networks* 45.3 (2004), p. 245–259.
- [255] Jiang-Jiang WANG, You-Yin JING, Chun-Fa ZHANG et Jun-Hong ZHAO. "Review on multi-criteria decision analysis aid in sustainable energy decision-making". In : *Renewable and Sustainable Energy Reviews* 13.9 (2009), p. 2263–2278.
- [256] Y.-S. WANG, N. MATNI et J. C. DOYLE. "Separable and Localized System Level Synthesis for Large-Scale Systems". In : *ArXiv e-prints* (jan. 2017). arXiv : [1701.05880 \[math.OC\]](https://arxiv.org/abs/1701.05880).
- [257] John Glen WARDROP. "Some theoretical aspects of road traffic research." In : *Proceedings of the institution of civil engineers* 1.3 (1952), p. 325–362.
- [258] Benjamin WATSON, M PASCAL, Oleg VERYOVKA, Andy FULLER, Peter WONKA et Chris SEXTON. "Procedural urban modeling in practice". In : *IEEE Computer Graphics and Applications* 3 (2008), p. 18–26.
- [259] Michael WEGENER et Franz FÜRST. "Land-use transport interaction : state of the art". In : *Available at SSRN* 1434678 (2004).
- [260] Michael WEGENER, Roger L MACKETT et David C SIMMONDS. "One city, three models : comparison of land-use/transport policy simulation models for Dortmund". In : *Transport Reviews* 11.2 (1991), p. 107–129.
- [261] Jörgen W WEIBULL. "An axiomatic approach to the measurement of accessibility". In : *Regional Science and Urban Economics* 6.4 (1976), p. 357–379.
- [262] JWR WHITEHAND, NJ MORTON et CMH CARR. "Urban morphogenesis at the microscale : how houses change". In : *Environment and Planning B : Planning and Design* 26.4 (1999), p. 503–515.
- [263] Norbert WIENER. *Cybernetics*. Hermann Paris, 1948.
- [264] Uri WILENSKY. "NetLogo". In : (1999).
- [265] Stephen WOLFRAM. *A new kind of science*. T. 5. Wolfram media Champaign, 2002.
- [266] Feng XIE et David LEVINSON. "How streetcars shaped suburbanization : a Granger causality analysis of land use and transit in the Twin Cities". In : *Journal of Economic Geography* (2009), lbp031.
- [267] Feng XIE et David LEVINSON. "Modeling the growth of transportation networks : A comprehensive review". In : *Networks and Spatial Economics* 9.3 (2009), p. 291–307.

- [268] Yihui XIE. "knitr : A general-purpose package for dynamic report generation in R". In : *R package version 1.7* (2013).
- [269] Kazuko YAMASAKI, Kaushik MATIA, Sergey V BULDYREV, Dongfeng Fu, Fabio PAMMOLLI, Massimo RICCABONI et H Eugene STANLEY. "Preferential attachment and growth dynamics in complex systems". In : *Physical Review E* 74.3 (2006), p. 035103.
- [270] Daniel YAMINS, Steen RASMUSSEN et David FOGEL. "Growing urban roads". In : *Networks and Spatial Economics* 3.1 (2003), p. 69–85.
- [271] Xin YE. "Investigation of Underlying Distributional Assumption in Nested Logit Model Using Copula-Based Simulation and Numerical Approximation". In : *Transportation Research Record : Journal of the Transportation Research Board* 2254 (2011), p. 36–43.
- [272] Bhanu M YERRA et David M LEVINSON. "The emergence of hierarchy in transportation networks". In : *The Annals of Regional Science* 39.3 (2005), p. 541–553.
- [273] Pierre ZEMBRI. "Les fondements de la remise en cause du Schéma Directeur des liaisons ferroviaires à grande vitesse : des faiblesses avant tout structurelles". In : *Annales de géographie*. JSTOR. 1997, p. 183–194.
- [274] Pierre ZEMBRI. "La contribution de la grande vitesse ferroviaire à l'interrégionalité en France.(High-speed rail and inter-regionality in France)". In : *Bulletin de l'Association de géographes français* 85.4 (2008), p. 443–460.
- [275] Kuilin ZHANG, Hani S MAHMASSANI et Chung-Cheng LU. "Dynamic pricing, heterogeneous users and perception error : Probit-based bi-criterion dynamic stochastic user equilibrium assignment". In : *Transportation Research Part C : Emerging Technologies* 27 (2013), p. 189–204.
- [276] Lei ZHANG et David LEVINSON. "The economics of transportation network growth". In : *Essays on transport economics*. Springer, 2007, p. 317–339.
- [277] Shanjiang ZHU et David LEVINSON. "Do people use the shortest path ? An empirical test of Wardrop's first principle". In : *91th annual meeting of the Transportation Research Board, Washington*. T. 8. Citeseer. 2010.

Cinquième partie

APPENDIX

10

AN INTERDISCIPLINARY APPROACH TO MORPHOGENESIS

This Appendix was submitted as an Essay Paper with C. Antelope (U. California), L. Hubatsch (F. Crick Institute) and J.M. Serna (Université Paris 7), as :

Antelope, C., Hubatsch, L., Raimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016.*

TECHNICAL DEVELOPMENTS

C'est hardcore tes calculs.

- ANONYME

This chapter gathers various technical developments, that have the common points to be not essential to the core of the thesis and difficult to digest.

11.1 DÉRIVATIONS POUR LES MODÈLES DE CROISSANCE URBAINE

Lemma 2 *The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.*

Proof Starting with first moment, we denote $\bar{P}_i(t) = \mathbb{E}[P_i(t)]$.

Independence of Gibrat growth rate yields directly

$\bar{P}_i(t) = \mathbb{E}[R_i(t)] \cdot \bar{P}_i(t-1)$. Starting for the preferential attachment model, we have $\bar{P}_i(t) = \mathbb{E}[P_i(t)] = \sum_{k=0}^{+\infty} k \mathbb{P}[P_i(t) = k]$. But

$$\{P_i(t) = k\} = \bigcup_{\delta=0}^{\infty} (\{P_i(t-1) = k - \delta\} \cap \{P_i \leftarrow P_i + 1\}^{\delta})$$

where the second event corresponds to city i being increased δ times between $t-1$ and t (note that events are empty for $\delta \geq k$). Thus, being careful on the conditional nature of preferential attachment formulation, stating that $\mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] = \lambda \cdot \frac{p}{P(t-1)}$ (total population $P(t)$ assumed deterministic), we obtain

$$\begin{aligned} \mathbb{P}[\{P_i \leftarrow P_i + 1\}] &= \sum_p \mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] \cdot \mathbb{P}[P_i(t-1) = p] \\ &= \sum_p \lambda \cdot \frac{p}{P(t-1)} \mathbb{P}[P_i(t-1) = p] = \lambda \cdot \frac{\bar{P}_i(t-1)}{P(t-1)} \end{aligned}$$

It gives therefore, knowing that $P(t-1) = P_0 + m \cdot (t-1)$ and denoting $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)}$

$$\begin{aligned}
\bar{P}_i(t) &= \sum_{k=0}^{\infty} \sum_{\delta=0}^{\infty} k \cdot \left(\lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \right)^{\delta} \cdot \mathbb{P}[P_i(t-1) = k - \delta] \\
&= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta') \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\
&= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot (\delta' + \bar{P}_i(t-1)) = \frac{q}{(1-q)^2} + \frac{\bar{P}_i(t-1)}{(1-q)} \\
&= \frac{\bar{P}_i(t-1)}{1-q} \left[1 + \frac{1}{\bar{P}_i(t-1)} \frac{q}{(1-q)} \right]
\end{aligned}$$

As it is not expected to have $\bar{P}_i(t) \ll P(t)$ (fat tail distributions), a limit can be taken only through λ . Taking $\lambda \ll 1$ yields, as

$0 < \bar{P}_i(t)/P(t) < 1$, that $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \ll 1$ and thus we can expand in first order of q , what gives

$$\bar{P}_i(t) = \bar{P}_i(t-1) \cdot \left[1 + \left(1 + \frac{1}{\bar{P}_i(t-1)} \right) q + o(q) \right]$$

$$\bar{P}_i(t) \simeq \left[1 + \frac{\lambda}{P_0 + m \cdot (t-1)} \right] \cdot \bar{P}_i(t-1)$$

It means that this limit is equivalent in expectancy to a Gibrat model with $\mu_i(t) = \mu(t) = 1 + \frac{\lambda}{P_0 + m \cdot (t-1)}$.

For the second moment, we can do an analog computation. We have still

$$\mathbb{E}[P_i(t)^2] = \mathbb{E}[R_i(t)^2] \cdot \mathbb{E}[P_i(t-1)^2]$$

and

$$\mathbb{E}[P_i(t)^2] = \sum_{k=0}^{+\infty} k^2 \mathbb{P}[P_i(t) = k]$$

We obtain the same way

$$\begin{aligned}
\mathbb{E}[P_i(t)^2] &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta')^2 \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\
&= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot \left(\mathbb{E}[P_i(t-1)^2] + 2\delta' \bar{P}_i(t-1) + \delta'^2 \right) \\
&= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} + \frac{2q\bar{P}_i(t-1)}{(1-q)^2} + \frac{q(q+1)}{(1-q)^3} \\
&= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} \left[1 + \frac{q}{\mathbb{E}[P_i(t-1)^2]} \left(\frac{2\bar{P}_i(t-1)}{1-q} + \frac{(1+q)}{(1-q)^2} \right) \right]
\end{aligned}$$

We have therefore an equivalence between the Gibrat model as a continuous formulation of a Preferential Attachment (or Simon model) in a certain limit. ■

11.2 SENSIBILITÉ DES LOIS D'ECHELLE URBAINES

We formalize the simple theoretical context in which we will derive the sensitivity of scaling to city definition. Let consider a polycentric city system, which spatial density distributions can be reasonably constructed as the superposition of monocentric fast-decreasing spatial kernels, such as an exponential mixture model [6]. Taking a geographical space as \mathbb{R}^2 , we take for any $\vec{x} \in \mathbb{R}^2$ the density of population as

$$d(\vec{x}) = \sum_{i=1}^N d_i(\vec{x}) = \sum_{i=1}^N d_i^0 \cdot \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|}{r_i}\right) \quad (10)$$

where r_i are spread parameters of kernels, d_i^0 densities at origins, \vec{x}_i positions of centers. We furthermore assume the following constraints :

1. To simplify, cities are monocentric, in the sense that for all $i \neq j$, we have $\|\vec{x}_i - \vec{x}_j\| \gg r_i$.
2. It allows to impose structural scaling in the urban system by the simple constraint on city populations P_i . One can compute by integration that $P_i = 2\pi d_i^0 r_i^2$, what gives by injection into the scaling hypothesis $\ln P_i = \ln P_{\max} - \alpha \ln i$, the following relation between parameters : $\ln [d_i^0 r_i^2] = K' - \alpha \ln i$.

To study scaling relations, we consider a random scalar spatial variable $a(\vec{x})$ representing one aspect of the city, that can be everything but has the dimension of a spatial density, such that the indicator $A(D) = \mathbb{E}[\iint_D a(\vec{x}) d\vec{x}]$ represents the expected quantity of a in area D . We make the assumption that $a \in \{0; 1\}$ ("counting" indicator) and that its law is given by $\mathbb{P}[a(\vec{x}) = 1] = f(d(\vec{x}))$. Following the empirical work done in [71], the integrated indicator on city i as a function of θ is given by

$$A_i(\theta) = A(D(\vec{x}_i, \theta))$$

where $D(\vec{x}_i, \theta)$ is the area centered in \vec{x}_i where $d(\vec{x}) > \theta$.

Assumption 1 ensures that the area are roughly disjoint circles. We take furthermore a simple amenity such that it follows a local scaling law in the sense that $f(d) = \lambda \cdot d^\beta$. It seems a reasonable assumption since it was shown that many urban variable follow a fractal behavior at the intra-urban scale [141] and that it implies necessarily a power-law distribution [62]. We make the additional assumption that $r_i = r_0$ does not depend on i , what is reasonable if the urban system is considered from a large scale. This assumption should be relaxed in numerical simulations. The estimated scaling exponent $\alpha(\theta)$ is then the result of the log-regression of $(A_i(\theta))_i$ against $(P_i(\theta))_i$ where $P_i(\theta) = \iint_{D(\vec{x}_i, \theta)} d$.

11.2.1 *Dérivation Analytique de la Sensibilité*

With above notations, let derive the expression of estimated exponent for quantity a as a function of density threshold parameter θ . The quantity computed for a given city i is, thanks to the monocentric assumption and in a spatial range and a range for θ such that $\theta \gg \sum_{j \neq i} d_j(\vec{x})$, allowing to approximate $d(\vec{x}) \simeq d_i(\vec{x})$ on $D(\vec{x}_i, \theta)$, is computed by

$$\begin{aligned} A_i(\theta) &= \lambda \cdot \iint_{D(\vec{x}_i, \theta)} d^\beta = 2\pi\lambda d_i^{0\beta} \int_{r=0}^{r_0 \ln \frac{d_i^0}{\theta}} r \exp\left(-\frac{r\beta}{r_0}\right) dr \\ &= \frac{2\pi d_i^{0\beta} r_0^2}{\beta^2} \left[1 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta - \left(\frac{\theta}{d_i^0} \right)^\beta \right] \end{aligned}$$

We obtain in a similar way the expression of $P_i(\theta)$

$$P_i(\theta) = 2\pi d_i^0 r_0^2 \left[1 + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0} - \frac{\theta}{d_i^0} \right]$$

The Ordinary-Least-Square estimation, solving the problem

$\inf_{\alpha, C} \|(\ln A_i(\theta) - C - \alpha \ln P_i(\theta))_i\|^2$, gives the value

$\alpha(\theta) = \frac{\text{Cov}[(\ln A_i(\theta))_i, (\ln P_i(\theta))_i]}{\text{Var}[(\ln P_i(\theta))_i]}$. As we work on city boundaries,

threshold is expected to be significantly smaller than center density,

i.e. $\theta/d_i^0 \ll 1$. We can develop the expression in the first order of

θ/d_i^0 and use the global scaling law for city sizes, what gives

$$\ln A_i(\theta) \simeq K_A - \alpha \ln i + (\beta - 1) \ln d_i^0 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta \text{ and}$$

$$\ln P_i(\theta) = K_P - \alpha \ln i + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0}. \text{ Developing the covariance and}$$

variance gives finally an expression of the scaling exponent as a

function of θ , where k_j, k_j' are constants obtained in the

development :

$$\alpha(\theta) = \frac{k_0 + k_1 \theta + k_2 \theta^\beta + k_3 \theta^{\beta+1} + k_4 \theta \ln \theta + k_5 \theta^\beta \ln \theta + k_6 \theta^\beta (\ln \theta)^2 + k_7 \theta^{\beta+1} (\ln \theta)^2}{k'_0 + k'_1 \ln \theta + k'_2 \theta \ln \theta + k'_3 \theta^2 + k'_4 \theta^2 \ln \theta + k'_5 \theta^2 (\ln \theta)^2} \quad (11)$$

This rational fraction predicts the evolution of the scaling exponent when the threshold varies. We study numerically its behavior in the next section, among other numerical experiments.

11.2.2 *Simulations*

Numériques

IMPLÉMENTATION We implement empirically the density model given in section 11.2. Centers are successively chosen such that in a given region of space only one kernel dominates in the sense that the sum of other contributions are above a given threshold θ_e . In

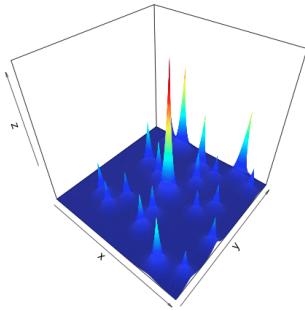


FIGURE 18 :

practice, adapting N to world size allows to respect the monocentric condition. Population are distributed in order to follow the scaling law with fixed α and r_i (arbitrary choice) by computing corresponding d_i^0 . Technical details of the implementation done in R [213] and using the package kernlab for efficient kernel mixture methods [139] are given as comments in source code¹. We show in figure 18 example of synthetic density distributions on which the numerical study is conducted. The validation of theoretical results on these experimental mixtures must still be conducted, along with sensitivity tests to random perturbations, influence of kernel type, and two-parameters phase diagram when adding in the computational model functional density distribution and associated cut-off threshold.

PERTURBATIONS ALÉATOIRES The simple model used is quite reducing for maximal densities and radius distribution. We aim to proceed to an empirical study of the influence of noise in the system by fixing d_i^0 and r_i the following way :

- d_i^0 follows a reversed log-normal distribution with maximal value being a realistic maximal density
- Radii are computed to respect rank-size law and then perturbed by a white noise.

TYPE DE NOYAU We shall test the influence of the type of spatial kernel used on results. We can test gaussian kernels and quadratic kernels with parameters within reasonable ranges analog to the exponential kernel.

¹ available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Scaling>

GENERATION OF CORRELATED SYNTHETIC DATA

APPLICATION : SÉRIES TEMPORELLES FINANCIÈRES

Contexte

Un premier domaine d'application proposé pour notre méthode est celui des séries temporelles financières, signaux typiques de systèmes complexes hétérogènes et multiscalaires [177] et pour lesquels les corrélations ont fait l'objet d'abondants travaux. Ainsi, l'application de la théorie des matrices aléatoires peut permettre de débruiter, ou du moins d'estimer la part de signal noyée dans le bruit, une matrice de correlations pour un grand nombre d'actifs échantillonnés à faible fréquence (retours journaliers par exemple) [41]. De même, l'analyse de réseaux complexes construits à partir des corrélations, selon des méthodes type arbre couvrant minimal [37] ou des extensions raffinées pour cette application précise [248], ont permis d'obtenir des résultats prometteurs, tels la reconstruction de la structure économique des secteurs d'activités. A haute fréquence, l'estimation précise de paramètres d'interdépendance dans le cadre d'hypothèses fixées sur la dynamique, fait l'objet d'importants travaux théoriques dans un but de raffinement des modèles et des estimateurs [17]. Les résultats théoriques doivent alors être testés sur des jeux de données synthétiques, qui permettent de contrôler un certain nombre de paramètres et de s'assurer qu'un effet prédit par la théorie est bien observable *toutes choses égales par ailleurs*. Par exemple, [199] dérive une correction du biais de l'estimateur de Hayashi-Yoshida qui est un estimateur de la covariance de deux browniens corrélés à haute fréquence dans le cas de temps d'observation asynchrones, par démonstration d'un théorème de la limite centrale pour un modèle généralisé endogénisant les temps d'observations. La confirmation empirique de l'amélioration de l'estimateur est alors obtenue sur un jeu de données synthétiques à un niveau de corrélation fixé.

Formalisation

CADRE Considérons un réseau d'actifs $(X_i(t))_{1 \leq i \leq N}$ échantillonnés à haute fréquence (typiquement 1s). On se place dans un cadre multi-scalaire (utilisé par exemple dans les approches par ondelettes [221] ou analyses multifractales du signal [42]) pour interpréter les signaux observés comme la superposition de

composantes à des multiples échelles temporelles : $X_i = \sum_{\omega} X_i^{\omega}$. On notera $T_i^{\omega} = \sum_{\omega' \leq \omega} X_i^{\omega'}$ le signal filtré à une fréquence ω donnée. Prédire l'évolution d'une composante à une échelle donnée est alors un problème caractéristique de l'étude des systèmes complexes, pour lequel l'enjeu est l'identification de régularités et leur distinction des composantes considérées comme stochastiques en comparaison¹. Dans un souci de simplicité, on représente un tel processus par un modèle de prédiction de tendance à une échelle temporelle ω_1 donnée, formellement un estimateur

GÉNÉRATION DES DONNÉES Il est alors aisé de générer \tilde{X}_i tel que $\text{Var}[\tilde{X}_i^{\omega_1}] = \Sigma R$ (Σ variance estimée et R matrice de corrélation fixée), par la simulation de processus de Wiener au niveau de corrélation fixé et tel que $X_i^{\omega \leq \omega_0} = \tilde{X}_i^{\omega \leq \omega_0}$ (critère de proximité au données : les composantes à plus basse fréquence qu'une fréquence fondamentale $\omega_0 < \omega_1$ sont identiques). En effet, si $dW_1 \perp\!\!\!\perp dW_1^\perp$ (et $\sigma_1 < \sigma_2$ pour fixer les idées, quitte à échanger les actifs), alors

$$W_2 = \rho_{12} W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2} \cdot \rho_{12}^2} W_1^{\perp\perp} \text{ est tel que } \rho(dW_1, dW_2) = \rho_{12}.$$

Les signaux suivants sont construits de la même manière par orthonormalisation de Gram. On isole alors la composante à la fréquence ω_1 voulue par filtrage, c'est à dire $\tilde{X}_i^{\omega_1} = W_i - \mathcal{F}_{\omega_0}[W_i]$ (avec \mathcal{F}_{ω_0} filtre passe-bas à fréquence de coupure ω_0), puis on reconstruit les signaux synthétiques par $\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1}$.

Implémentation

et

résultats

MÉTHODOLOGIE La méthode est testée sur un exemple de deux actifs du marché des devises (EUR/USD et EUR/GBP), sur une période de 6 mois de juin 2015 à novembre 2015. Le nettoyage des données², originellement échantillonnées à l'ordre de la seconde, consiste dans un premier temps à la détermination du support temporel commun maximal (les séquences manquantes étant alors

¹ voir [107] pour une discussion étendue sur la construction de *schema* pour l'étude de systèmes complexes adaptatifs (par des systèmes complexes adaptatifs).

² obtenues depuis <http://www.histdata.com/>, sans licence spécifiée, les données nettoyées et filtrées à ω_m uniquement sont mises en accessibilité pour respect du copyright.

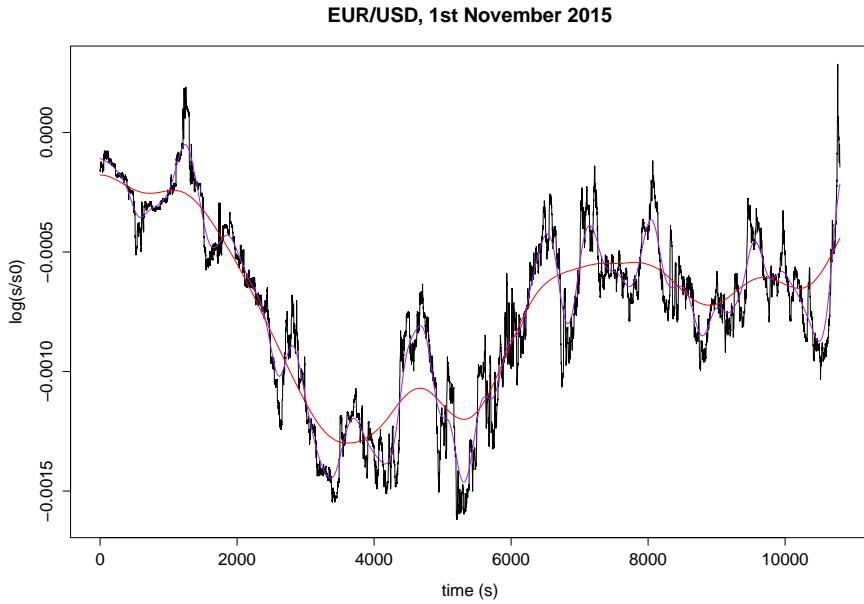


FIGURE 19 :

ignorées, par translation verticale des séries, i.e. $S(t) := S(t) \cdot \frac{S(t_n)}{S(t_{n-1})}$ lorsque t_{n-1}, t_n sont les extrémités du “trou” et $S(t)$ la valeur de l’actif, ce qui revient à garder la contrainte d’avoir des retours à pas de temps similaires entre actifs). On étudie alors les *log-prix* et *log-retours*, définis par $X(t) := \log \frac{S(t)}{S_0}$ et $\Delta X(t) = X(t) - X(t-1)$. Les données brutes sont filtrées à une fréquence $\omega_m = 10\text{min}$ (qui sera la fréquence maximale d’étude) pour un souci de performance computationnelle. On utilise un filtre gaussien non causal de largeur totale ω . On fixe $\omega_0 = 24\text{h}$ et on se propose de construire des données synthétiques aux fréquences $\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$. Voir la figure 19 pour un exemple de la structure du signal à ces différentes échelles.

Il est crucial de noter l’interférence entre les fréquences ω_0 et ω_1 dans le signal construit : la corrélation effectivement estimée est

$$\rho_e = \rho [\Delta \tilde{X}_1, \Delta \tilde{X}_2] = \rho [\Delta T_1^{\omega_0} + \Delta \tilde{X}_1^\omega, \Delta T_2^{\omega_0} + \Delta \tilde{X}_2^\omega]$$

ce qui conduit à dériver dans la limite raisonnable $\sigma_1 \gg \sigma_0$ (fréquence fondamentale suffisamment basse), lorsque $\text{Cov} [\Delta \tilde{X}_i^{\omega_1}, \Delta X_j^\omega] = 0$ pour tous $i, j, \omega_1 > \omega$, et les retours d’espérance nulle à toutes échelles, en notant $\rho_0 = \rho [\Delta T_1^{\omega_0}, \Delta T_2^{\omega_0}]$, $\rho = \rho [\tilde{X}_1^{\omega_1}, \tilde{X}_2^{\omega_1}]$, et $\epsilon_i = \frac{\sigma(\Delta T_i^{\omega_0})}{\sigma(\Delta \tilde{X}_i^{\omega_1})}$, la correction sur la corrélation

effective due aux interférences : la correlation effective est alors au premier ordre

$$\rho_e = [\varepsilon_1 \varepsilon_2 \rho_0 + \rho] \cdot \left[1 - \frac{1}{2} (\varepsilon_1^2 + \varepsilon_2^2) \right] \quad (12)$$

ce qui donne l'expression de la correlation que l'on pourra effectivement simuler dans les données synthétiques.

La correlation est estimée par méthode de Pearson, avec l'estimateur de la covariance au biais corrigé, c'est à dire

$$\hat{\rho}[X_1, X_2] = \frac{\hat{C}[X_1, X_2]}{\sqrt{\hat{V}\text{ar}[X_1]\hat{V}\text{ar}[X_2]}}, \text{ où}$$

$$\hat{C}[X_1, X_2] = \frac{1}{(T-1)} \sum_t X_1(t)X_2(t) - \frac{1}{T(T-1)} \sum_t X_1(t) \sum_t X_2(t) \text{ et}$$

$$\hat{V}\text{ar}[X] = \frac{1}{T} \sum_t X^2(t) - \left(\frac{1}{T} \sum_t X(t) \right)^2.$$

Le modèle de prédiction M_{ω_1} testé est simplement un modèle ARMA pour lequel on fixe les paramètres $p = 2$, $q = 0$ (on ne crée pas de correlation retardée, on ne s'attend donc pas à de grand ordre d'auto-regression, les signaux originaux étant à mémoire relativement courte ; de plus le lissage n'est pas nécessaire puisqu'on travaille sur des données filtrées), appliqué de manière adaptative³. Plus précisément, étant donné une fenêtre temporelle T_W , on estime pour tout t le modèle sur $[t - T_W + 1, t]$ afin de prédire les signaux à $t + 1$.

IMPLÉMENTATION L'implémentation est faite en language R, utilisant en particulier la bibliothèque MTS [247] pour les modèles de séries temporelles. Les données nettoyées et le code source sont disponibles de manière ouverte sur le dépôt git du projet⁴.

RÉSULTATS La figure ?? donne les correlations effectives calculées sur les données synthétiques. Pour des valeurs standard des paramètres (par exemple pour $\omega_0 = 24h$, $\omega_1 = 2h$ et $\rho = -0.5$), on a $\rho_0 \simeq 0.71$ et $\varepsilon_i \simeq 0.3$ et ainsi $|\rho_e - \rho| \simeq 0.05$. On constate dans l'intervalle $\rho \in [-0.5, 0.5]$ un bon accord entre la valeur ρ_e prédite par 12 et les valeurs observées, et une déviation pour de plus grandes valeurs absolues, d'autant plus grande que ω_1 est petit : cela confirme l'intuition que lorsque la fréquence descend et se rapproche de ω_0 , les interférences entre les deux composantes vont devenir non négligeables et invalider les hypothèses d'indépendance par exemple.

³ il s'agit d'un niveau d'adaptation relativement faible, les paramètres T_W, p, q et même le type de modèle restant fixés. On se place ainsi dans le cadre de [198] qui suppose une dynamique localement paramétrique, mais pour lequel on fixe les métaparamètres de la dynamique. On pourrait imaginer estimer un T_W variable qui s'adapterait pour une meilleure estimation locale, à l'image de l'estimation de paramètres en traitement du signal Bayesien effectuée via augmentation de l'état par les paramètres.

⁴ at <https://github.com/JusteRaimbault/SynthAsset>

On applique ensuite le modèle prédictif décrit ci-dessus aux données synthétiques, afin d'étudier sa performance moyenne en fonction du niveau de correlation des données. Les résultats pour $\omega_1 = 1\text{h}, 1\text{h}30, 2\text{h}$ sont présentés en figure 21. Le résultat a priori contre-intuitif d'une performance maximale à correlation nulle pour l'un des actifs confirme l'intérêt d'une génération de données hybrides : l'étude des correlations décalées (*lagged correlations*) montre une dissymétrie présente dans les données réelles, interprété à l'échelle journalière comme une influence augmentée de EURGBP sur EURUSD à 2h de décalage environ. L'existence de ce *lag* permet une "bonne" prédiction de EURUSD due à la fréquence fondamentale, perturbée par le bruit ajouté, de façon proportionnelle à sa corrélation : plus les bruits sont corrélés, plus le modèle les prendra en compte et se trompera plus à cause du caractère markovien des browniens simulés⁵.

L'exemple présenté ici est un *modèle jouet* et n'a pas d'application pratique, mais démontre l'intérêt de l'utilisation des données synthétiques simulées. On peut imaginer simuler des données plus proches de la réalité (existence de motifs réalistes de *lagged correlation* par exemple, modèles plus réalistes que le Black-Scholes) et appliquer la méthode sur des modèles plus opérationnels.

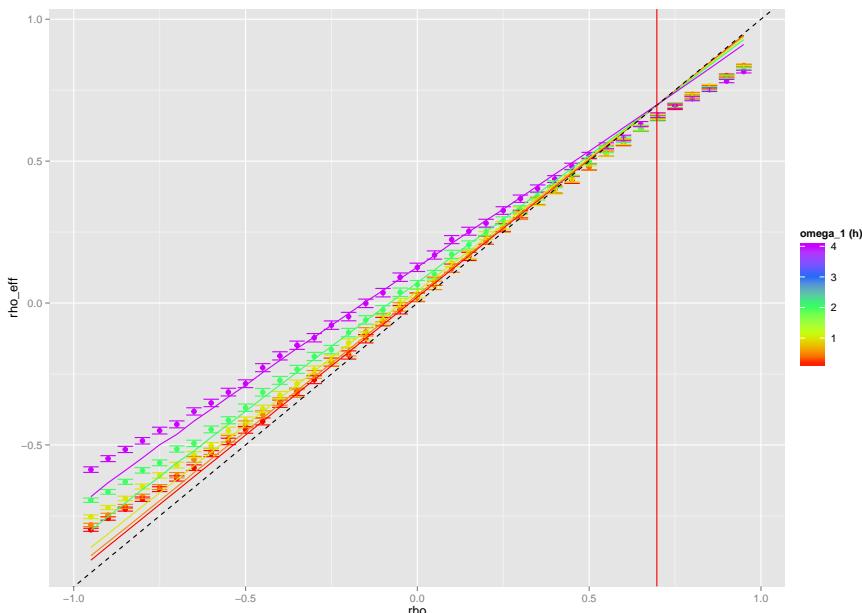


FIGURE 20 :

⁵ en théorie le modèle utilisé n'a aucun pouvoir prédictif sur des browniens purs

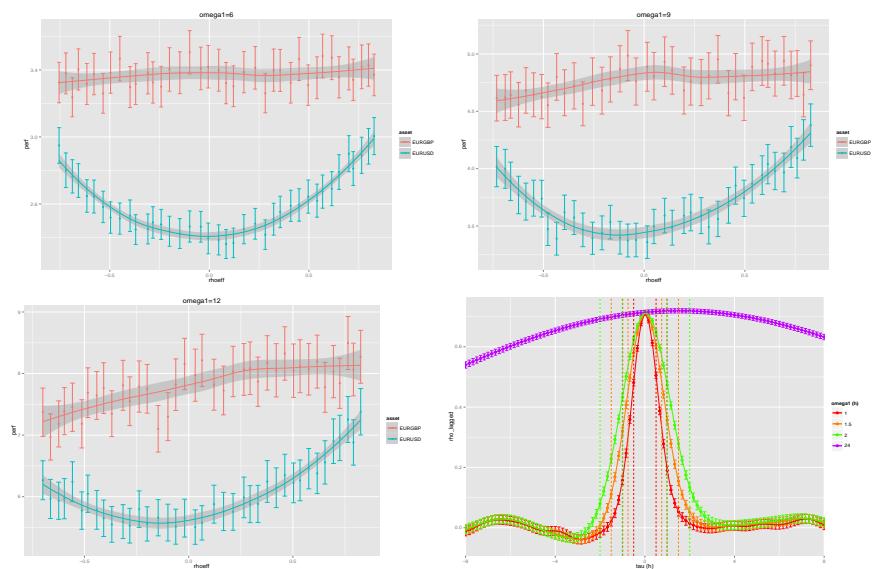


FIGURE 21 :

14

DATASETS

This appendix lists and describes the different open datasets created and used in the thesis.

14.1 DONNÉES DE TRAFFIC DU GRAND PARIS

14.2 RÉSEAU ROUTIER EUROPÉEN

14.3 RÉSEAU DYNAMIQUE DES AUTOROUTES FRANÇAISES

15

SOFTWARES AND PACKAGES

This appendix lists and describes the different open datasets created and used in the thesis.

15.1 LARGENETWORK : IMPORT DE RÉSEAU ET SIMPLIFICATION
POUR R

15.2 FOUILLE DE CORPUS SCIENTIFIQUE

ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION

*You must not be afraid of putting
code in your thesis, code is not dirty*
- ALEXIS DROGOUL

And yet it is. It makes no sense to put code listings in the core of the text if there is no particular algorithmic detail that requires attention. As soon as implementation biases are avoided, architecture and source for a computational model should be independent from its formal description (but provided along model description with source code as already mentioned before). We give in this appendix architectural details on main models of simulation or algorithms we used. Langage and size (in code lines) are provided, along with architectural remarkable features. See <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models>

for all models, empirical analysis and small experiments. The following reports are partially generated automatically using experimental tools aimed at workflow improvement.

16.1 REVUE

SYSTÉMATIQUE

ALGORITHMIQUE

OBJECTIFS Implement systematic literature review algorithm.

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp>

CARACTÉRISTIQUES

- Language : Java
- Size : 7116

PARTICULARITÉS

- HashConsing used for unique bibliography object, specific hashcode switching if id available or only titles (proceed to lexical distance comparison in that latest case).
- API to cortex currently being replaced by Python scripts.

ARCHITECTURE Classical object oriented, see code.

SCRIPTS ADDITIONNELS R for result exploration and visualization.

16.2 BIBLIOMÉTRIE

INDIRECTE

OBJECTIFS Hypernetworks analysis of cybergeo journal.

LOCALISATION <https://github.com/Geographie-cites/cybergeo20/tree/master/HyperNetwork>
<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp> for common Java part.

CARACTÉRISTIQUES

- Language : Python, R and Java.
- Size : -

PARTICULARITÉS Polyglot

ARCHITECTURE See schema chapter 3.

SCRIPTS ADDITIONNELS -

16.3 CROISSANCE

URBAINE

OBJECTIF Simple density urban growth model.

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Density>

CARACTÉRISTIQUES

- Language : NetLogo then scala.
- Size : 4355

PARTICULARITÉS Morphological indicators in scala implemented with Fast Fourier transform ; with R communication in NetLogo.

ARCHITECTURE Nothing particular.

SCRIPTS ADDITIONNELS R for result exploration and morphological analysis.
 oms for model exploration.

16.4 GÉNÉRATION DES DONNÉES SYNTHÉTIQUES CORRÉLÉES

OBJECTIFS Weak coupling of density generation and network generation.

LOCALISATION https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Network_20151229

CARACTÉRISTIQUES

- Language : NetLogo (network) and scala.
- Size : 3188

PARTICULARITÉS Network heuristic easier to implement and explore in netlogo

ARCHITECTURE OpenMole allows coupling between modules through exploration script.

SCRIPTS ADDITIONNELS R for result exploration.
oms for model exploration.

16.5 MODÈLE

LUTECIA

OBJECTIF Implementation of Lutecia model, chapter 6.

LOCALISATION

<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Governance/MetropolSim/Lutecia>

CARACTÉRISTIQUES

- Language : NetLogo
- Size : 4791

PARTICULARITÉS Shortest path dynamical programming using matrices.

ARCHITECTURE Pseudo object architecture in agent environment.

SCRIPTS ADDITIONNELS R for result exploration.
oms for model exploration.

16.6 ANALYSE

DES

RÉSEAUX

OBJECTIF Simplification of european road network

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/StaticCorrelations>

CARACTÉRISTIQUES

- Language : R, Shell, PostgreSQL
- Size : 505

PARTICULARITÉS Handling of large size databases imposes sequential processing ; use of external program osmosis for conversion from osm data to pgsql.

ARCHITECTURE Shell script lead maneuvers.

SCRIPTS ADDITIONNELS -

TOOLS AND WORKFLOW FOR AN OPEN REPRODUCIBLE RESEARCH

Open for Discovery
- PLoS

We briefly evoke here tools or workflows currently under development or testing, aimed at easing an open reproducible research and making it more transparent.

17.1 GÉNÉRATEUR DE DOCUMENTATION NETLOGO

Documentation generation is central for reproducibility as it can automatize implementation description. NetLogo does not provide a documentation generator and we are thus currently writing a Doxygen wrapper for NetLogo code, that basically consists in transforming NetLogo code into Java code and parsing documentation comment blocks. An experimental version is available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Doc>.

17.2 GIT COMME OUTIL DE REPRODUCTIBILITÉ

The use if git as a reproducibility and transparency tool was emphasized in [220] (for various reasons such as exact history tracing, easy cloning, past commit branching). It furthermore can help individual workflow for advantages such as automatic backup, organisation, experiments tracking. We use it actively and develop extensions for it.

17.3 GIT-DATA

git-data is a shell based (experimental) git extension, available at <https://github.com/JusteRaimbault/gitdata>, that allows automatized backup of large file within a git repository, their transparent integration in ignored files and the creation of symbolic links for a transparent local use.

17.4 VERS UN GESTIONNAIRE DE MÉTADONNÉES COMPATIBLE AVEC GIT

The issue of meta-data for figures is a crucial issue, as it is often difficult to keep a trace of all parameter values that have generated it, along with the corresponding code. Tricks may furthermore happen in script environments such as R or python when variables are accidentally modified without code modification. Keeping an exhaustive trace of the exact dataset, code and history that has generated a precise figure is a necessary condition for exact reproducibility. We are elaborating a git-compatible tool that would automatically handle these metadata, for example by branching and associating the unique commit hash to the figure. To become not an organizational burden nor a repository perturbation, we must still make some experiments. The final idea would be to have under each figure a unique identifier linking to the associated reproducing environment.

17.5 TORPOOL

TorPool is a java based Tor wrapper available with an api (currently only java, R version projected) at <https://github.com/JusteRaimbault/TorPool>. It allows among other purposes tricky data retrieval.